Declaring and Diagnosing Research Designs[†]

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

First draft: 5/7/2016
This draft: 8/15/2018

Abstract

Researchers need to select high-quality research designs and communicate those designs clearly to readers. Both tasks are difficult. We provide a framework for formally "declaring" the analytically relevant features of a research design in a demonstrably complete manner, with applications to qualitative, quantitative, and mixed methods research. The approach to design declaration 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, correct identification of causal conditions, and other "diagnosands." Ex ante declarations can be used to improve designs and facilitate preregistration, analysis, and reconciliation of intended and actual analyses. Ex post declarations are 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, Tom Leavitt, Winston Lin, Matto Mildenberger, Matthias Orlowski, Molly Roberts, 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. We thank Tara Slough for helping in concept development. We thank Clara Correia and Lily Medina for excellent research support and Neal Fultz, Luke Sonnet, Clara Correia, Lily Medina, and Aaron Rudkin for their collaboration on software development. The methods proposed in this paper are implemented in an accompanying open-source software package, `DeclareDesign` (Blair et al. 2018).

[‡]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`

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

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

To select strong designs, we often rely on rules of thumb, simple power calculators, or principles from the methodological literature that typically address one component of a design while assuming optimal conditions for others. These relatively informal practices can 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 often defend them with references to previous studies that used similar approaches; with power analyses that may rely on assumptions unknown even to ourselves; or, in rare cases, using *ad hoc* simulation code. In cases where there is a dispute over the merits of different approaches, disagreements sometimes fall back to first principles or epistemological debates rather than on demonstrations of the conditions under which one approach does better than another.

In this paper we describe an approach to address these problems. We introduce a framework—*MIDA*—that asks researchers to specify 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 that help to evaluate it. Familiar diagnosands include the power of a design, the bias of an estimator with respect to an estimand, or the coverage probability of a procedure for generating confidence intervals. We say a design declaration is "diagnosand-complete" when a diagnosand can be estimated from the declaration. 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 diagnosands are valuable and in turn what features must be declared. In practice, domain-specific standards might be agreed upon among members of particular research communities. For instance, researchers 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, or for a diagnosand directly related to policy choices, such as the probability of making the right policy decision after research is conducted.

Using this framework researchers can *declare* research designs as computer code objects and

then *diagnose* the statistical properties of the design relying on this declaration. We view declaring and diagnosing designs as a distinct step in the research process, separate from the familiar tasks of theory-building, obtaining approvals, pre-registering analysis protocols (if applicable), collecting data, presenting results to others, and preparing the final write-up for peer-review. We emphasize that the term "declare" does not imply a public declaration or even necessarily 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. Researchers can declare and diagnose their designs with the companion software for this paper, `DeclareDesign`, but the principles of design declaration and diagnosis do not depend on any particular software implementation.

The formal characterization and diagnosis of designs before implementation can serve many purposes. First, researchers can learn about and improve their inferential strategies. Done at this stage, diagnosis of a design and alternatives can help a researcher select from a range of designs, conditional upon beliefs about the world. Later, a researcher may include design declaration and diagnosis as part of a preanalysis plan or in a funding request. At this stage the full specification of a design serves a communication function and enables third parties to understand a design and an author's intentions. Even if 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.

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 more inductive research, 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 usefully 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 been realized. This may be particularly true of interpretavist approaches and approaches to process tracing that work backwards from outcomes to a set of possible causes that cannot be prespecified. We acknowledge from the outset that variation in research strategy limits the utility of our procedure

for different types of research.

Even still, as we highlight below, the reach of our framework appears great and can include designs that include a focus on discovery, qualitative inference, and different approaches to mixed methods research, as well as designs that focus on "effects-of-causes" questions, often associated with quantitative approaches, and "causes-of-effects" questions, often associated with qualitative approaches.

Formally declaring research designs as objects in the manner we describe here brings, we hope, four benefits. It can facilitate the diagnosis of designs in terms of their ability to perform under specified conditions; it can assist in the improvement of research designs through comparison with alternatives; it can enhances research transparency by making design choices explicit; and it can provide strategies to assist principled replication and reanalysis of published research.

1. Research Designs and Diagnosands

We present a general description of a research design as the specification of a problem and a strategy to answer it. We build on two influential research design frameworks. King, Keohane and Verba (1994, p. 13) enumerate four components of a research design: a theory, a research question, data, and an approach to using the data. Geddes (2003) articulates the links between theory formation, research question formulation, case selection and coding strategies, and strategies for case comparison and inference. In both cases the set of steps are closely aligned to those in the framework we propose. In the exposition, we also employ elements from Pearl's (2009) approach to structural modeling, which provides a syntax for mapping design inputs to design outputs as well as the potential outcomes framework as presented, for example, in Imbens and Rubin (2015), which many political scientists use to clarify their inferential targets. We characterize the design problem at a high level of generality with the central focus being on the relationship between questions and answer strategies. We further situate the framework within existing literature in section 2.

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, as in Imbens and Rubin (2015), "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. 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: $\Pr(X_1|do(X_2 = 1))$.[1] We let a^M denote the answer to I *under the model*. Conditional on the model, a^M is the value of the estimand, the quantity that the researcher wants to learn about.

3. A **data** strategy, D, generates data d on X. Data d arises, under model M with probability $P_M(d|D)$. The data strategy includes sampling strategies and assignment strategies, which we denote with P_S and P_Z respectively. Measurement techniques are also a part of data strategies and can be thought of as a selection of observable variables that carry information about unobservable variables.

4. An **answer** strategy, A, that generates answer a^A using data d.

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 $P_M(a^M -$

[1] 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.

$a^A|D$). 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.

MIDA captures the analysis-relevant features of a design, but it does not describe substantive elements, such as how theories are derived or interventions are implemented. 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. However, information in *MIDA* is typically insufficient to assess those substantive elements, an important and separate part of assessing the quality of a research study.

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 a given strategy? Would you do better, given some desideratum, 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" depends on the model (since the ATE depends on the model's assumptions about potential outcomes). 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 are governed by probability distributions that arise because 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 and Z (all defined on the reals). These form the signature. In additional we assume 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 inquiry I is "what would be the average effect of a unit increase 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.

We have specified all the components of MIDA. We now ask: How strong is this 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, as measured by this diagnosand, because the data and analysis strategy do not handle the confounding described by the model (see Online Appendix Section S1 for a formal declaration and diagnosis of this design). 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 the underlying process involved nonlinearities).

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

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 a decision based on estimated effect yields net benefits	✓	✓	✓	✓
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.

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.

The set listed here includes many canonical diagnosands used in classical quantitative analyses. 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. One might conclude that an intervention gives "value for money" if estimates are of a certain size and be interested in the probability that a researcher in correct in concluding that an intervention provides value for money.

Some of these diagnosands apply to qualitative research strategies also, such as Type *S* error rate or value for money, but some, such as statistical power, clearly do not. We believe there is not yet a consensus around diagnosands for qualitative designs though in certain treatments clear analogues of diagnosands exist, such as power, coverage and consistency for QCA researchers or correct identification of causes of effects, or of causal pathways, for scholars using process

tracing.

Though many of these diagnosands are familiar to scholars using frequentist approaches, analogous diagnosands can be used to assess Bayesian estimation strategies (see Rubin 1984), and as we illustrate below, some diagnosands are unique to Bayesian strategies.

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.[2]

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 a 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 in Table 1). In contrast, Bias- or RMSE-completeness does not require a hypothesis test, but does require the specification of an estimand.

Diagnosand-completeness is a desirable property to the extent that it means a diagnosand can be calculated. How useful this is depends however on how useful the diagnosand is for decision making. Thus evaluating completeness should focus first on whether diagnosands for which completeness holds are indeed useful ones.

This usefulness depends in part on whether the information on which diagnoses are made is *believable*. A design may be bias-complete for instance under the assumptions of a particular spillover structure, for example. Readers may disagree with these assumptions but there are still gains from the declaration as the grounds for claims for unbiasedness are clear and the effects of deviations from model assumptions can be assessed. In practice, different research communities set different standards for what constitutes sufficient information to make such conjectures about

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

the world plausible.

2. Existing Approaches to Learning About Research Designs

The MIDA framework can be distinguished from much quantitative research design advice by its focus on the ways in which multiple components of a research design relate to each other. Statistics articles and textbooks tend to focus on a specific class of estimators (Angrist and Pischke 2008; Rosenbaum 2002; Imbens and Rubin 2015), set of estimands (Heckman, Urzua and Vytlacil 2006; Imai, King and Stuart 2008; Deaton 2009; Imbens 2010), data collection strategies (Lohr 2010), or ways of thinking about data-generation models (Gelman and Hill 2006; Pearl 2009). In Shadish, Cook and Campbell (2002, p 156), for example, the "elements of a design" consist of assignment, measurement, comparison groups and treatments," a definition that does not include questions of interest or estimation strategies.[3]

In contrast, a number of qualitative treatments focus more on the many stages of a research design, from theory generation, to case selection, measurement, and inference. In an influential book on mixed method research design for comparative politics, for example, Geddes (2003) articulates the links between theory formation (M), research question formulation (I), case selection and coding strategies (D), and strategies for case comparison and inference (A). King, Keohane and Verba (1994) and the ensuing discussion in Brady and Collier (2010) highlight how alternative qualitative strategies present tradeoffs in terms of diagnosands such as bias and generalizability. But few of these texts investigate those diagnosands formally in order to measure the size of the tradeoffs between alternative qualitative strategies.[4]

Indeed, far from being fragmented, qualitative approaches, including process tracing and qualitative comparative analysis, sometimes appear almost hermetic, complete with specific epistemologies, types of research questions, modes of data gathering, and analysis. Though integrated, these strategies are not generally formalized and the necessity for the different components to go together, as an approach, can be hard to demonstrate.

[3] In some instances, quantitative researchers do present multiple elements of research design, though in a more fragmented way than in MIDA: Gerber and Green (2012), for example, examine data-generating models, estimands, assignment and sampling strategies, and estimators for use in experimental causal inference; and Shadish, Cook and Campbell (2002) and Dunning (2012) similarly describe the various aspects of designing quasi-experimental research and exploiting natural experiments.

[4] An exception is provided by Herron and Quinn (2016), who conduct a formal investigation of the RMSE and bias exhibited by the alternative case selection strategies proposed in an influential piece by Seawright and Gerring (2008).

The relatively fragmented manner in which the quantitative design is thought of in the literature may produce some real research risks for individual research projects. In contrast, the more holistic approaches of some qualitative traditions can make interaction with more statistical approaches difficult. We expand on these points next.

A useful way to illustrate the fragmented nature of thinking on research design among quantitative scholars is to examine the tools that are actually used to do research design. Perhaps the most prominent of these are "power calculators" which have an all-design flavor in the sense that they ask whether given an analysis strategy, a data collection strategy is likely to be able to answer a particular question. Power calculations like these are done using formulae (e.g., Cohen 1977; Haseman 1978; Muller and Peterson 1984; Muller et al. 1992; Lenth 2001); software tools such as Web applications and general statistical software (e.g., `easypower` for `R` and Power and Sample Size for Stata) and standalone tools (e.g. Optimal Design, G*Power, nQuery, SPSS Sample Power); and sometimes Monte Carlo simulations.

In most cases these tools, though touching on multiple parts of a design, in fact leave almost no scope to describe what the data generating processes can be, what the questions of interest are, and what types of analyses will be conducted. We conducted a census of currently available computational diagnostic tools and assessed their ability to correctly diagnose three variants of a very common experimental design, in which assignment probabilities are heterogeneous by block.[5] The first variant simply uses a difference-in-means estimator (DIM), the second conditions on block fixed effects (BFE), and the third includes inverse-probability weighting to account for the heterogeneous assignment odds (BFE-IPW).

We found that the vast majority of tools used are unable to correctly characterize the tradeoffs these three variants present. As shown in Table 2, none of the tools was able to diagnose the design while taking account of important features that bias unweighted estimators,[6] and therefore they tend to exaggerate the study's power by about fourteen percentage points.

[5]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." We found no admissible tools using the term "statistical bias calculator." Thirty of the 143 tools we identified were able to diagnose inferential properties of designs, such as their power. See Online Appendix Section S3 for further details on the tool survey.

[6]For example, no design could account for: 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. 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. Results correspond to design tool census concluded in July 2017 and do not include tools published since then.

Because such tools typically only require information on M (and occasionally on A), none was able to calculate the power for the IPW-BFE estimator. Moreover, no tool sought to calculate the design's bias, root mean-squared-error, or coverage. The companion software to this article, which was designed based on MIDA, illustrates that power is a misleading indicator of quality in this context, however. While the IPW-BFE estimator is better powered and less biased (in terms of the PATE) than the BFE estimator, its purported efficiency is misleading. IPW-BFE is better powered than DIM and BFE 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.

We draw a number of conclusions from this review of tools.

First researchers are generally not designing studies using the actual strategies that they will use to conduct analysis. Looked at from the perspective of the overall designs, the power calculations are answering the wrong questions.

Second, the tools can drive scholars towards relatively narrow design choices. The levers that seem available given the focus on power focus primarily on parts of a data strategy (sample size, number of clusters) and, more curiously, on effect sizes—that is, the estimands themselves, which researchers might not want or be able to alter. But they do not generally focus on broader aspects of a data strategy (such as different assignment strategies) or on answer strategies (such as choice of estimator). While researchers may have an awareness that such tradeoffs exist, quantifying the *extent* of the tradeoff is by no means obvious until one declares the model, inquiry, data strategy

and answer strategies in code.

Third, the tools focus attention on a relatively narrow set of questions for evaluating a design. While understanding power is important for some designs, the range of possible diagnosands of interest is much broader. Quantitative researchers tend to focus on power to the detriment of other diagnosands such as bias, coverage, or root mean square error. MIDA makes clear, however, that these features of a design are often linked in ways that current practice obscures.

A second illustration of risks arising from a fragmented conceptualization of the elements of a research design comes from debates over the focus of the identification revolution on properties of estimators without equal consideration to the estimands being estimated. Huber (2013) for example worries that the focus on identification leads researchers away from asking compelling questions. In the extreme, the estimators themselves (and not the researchers) appear to select the estimand of interest. Thus Deaton (2009) highlights how instrumental variables approaches identify effects for a subpopulation of compliers. The identity of this subpopulation may not be known. Moreover, there may be little reason why this subpopulation may be of particular interest. Indeed as researchers swap one instrument for another the implicit estimand changes. Deaton's worry is that researchers are getting an answer, but they do not know what the question is. Aronow and Samii (2016) express a similar concern for models using regression with controls. Were the question posed as the average effect of a treatment, then the performance of the instrument would depend on how well the instrumental variables estimate estimates that quantity, and not how well they answer the question for a different subpopulation. This is not done in usual practice however as estimands are often not included as part of a research design.

Finally, the combination of a fractured approach to design in the formal quantitative literature, and the holistic but often less formal approaches in the qualitative literature may limit the ability of these approaches to learn from each other.

Goertz and Mahoney (2012) tell a tale of two cultures in which qualitative and quantitative researchers differ not just in the analytic tools they use, but in very many ways, including, fundamentally, in their conceptualizations of causation and the kinds of questions they ask. They highlight how qualitative researchers think of causation in terms of necessary and/or sufficient causes whereas many quantitative researchers focus on potential outcomes and average effects. One might worry that such differences would preclude design declaration within a common

framework. Yet, although these differences in orientations exist, this difference in representation does not imply differences in the formal definition of any given estimand, at least for qualitative scholars that consider causes in counterfactual terms.[7]

For example a representation of a causal process in terms of causal configurations might take the form: $Y = AB + C$, meaning that the presence of A and B or the presence of C is sufficient to produce Y. This configuration statement maps directly into a potential outcomes function (or structural equation) of the form $Y(A, B, C) = \max(AB, C)$. Given this, the marginal effect of one variable, conditional on others, can be translated to the conditions in which the variable is difference making in the sense of altering relevant INUS conditions: $E(Y(A = 1|B, C) - Y(A = 0|B, C)) = E(B = 1, C = 0)$.[8] Describing these differences in notation as differences in notions of causality suggests that there is limited scope for considering designs that mix approaches, and that there is little that practitioners of one approach can say to practitioners of another approach. In contrast clarification that the difference is one regarding the inquiry—i.e., which combinations of variables guarantee a given outcome and not the average marginal effect of a variable across conditions—opens up the possibility to assess how quantitative estimation strategies fare when applied to estimating this estimand.

A second point of difference is nicely summarized by Goertz and Mahoney (2012, p. 230): "qualitative analysts adopt a 'causes-of-effects' approach to explanation [...] statistical researchers follow the 'effects-of-causes' approach employed in experimental research." We agree with this association though from a *MIDA* perspective we see such distinctions as differences in estimands and not as differences in ontology. Conditioning on a given X and a Y the effects of cause question is $E(Y(X = 1) - Y(X = 0))$ and the cause of effects question is $1 - E(Y(X = 0)|Y(X = 1) = 1)$ (with similar expressions when there are multiple explanatory variables). The two questions are of a similar form though the causes of effects question is harder to answer (Dawid 2000). Once thought of as questions about what the estimand is one can assess directly when one or other estimation strategy is more or less effective at learning about the estimand of interest. In fact

[7]Schneider and Wagemann (2012, pg. 320-1) also note that there are not grounds to assume incommensurability, noting that "if set-theoretic, method-specific concepts.... can be translated into the potential outcomes framework, the communication between scholars from different research traditions will be facilitated." See also Mahoney (2008) on the consistency of these conceptualizations.

[8]Goertz and Mahoney (2012, pg.59) also make the point that the difference is in practice, and is not fundamental: "Within quantitative research, it does not seem useful to group cases according to common causal configurations on the independent variables. *Although one could do this*, it is not a practice within the tradition." (Emphasis added.)

experiments are in general not able to solve the identification problem for causes of effects questions (Dawid 2000) and this may be one poor reason for why these questions are often ignored. Exceptions include Yamamoto (2012) and Balke and Pearl (1994).

Below we demonstrate gains from declaration of designs in a common framework by providing examples of design declaration for crisp-set qualitative comparative analysis (Ragin 1987), nested case analysis (Lieberman 2005), and CPO (causal process observation) process-tracing (Collier 2011; Fairfield 2013), alongside experimental and quasi experimental designs.

Overall this discussion suggests that the common ways in which designs are conceptualized produce three distinct problems. First, in practice the different components of a design may not be chosen to work optimally together. Second, improper weight is allocated across components of a design. Third, the absence of a common framework across research traditions obscures where the points of overlap and difference lie and may limit both critical assessment of approaches and cross fertilization over approaches. We hope that the *MIDA* framework and tools can help address these challenges.

3. Declaring and Diagnosing Research Designs in Practice

A design that can be declared in computer code can then be simulated in order to diagnose its properties. The approach to declaration that we advocate is one that conceives of a design as a concatenation of steps. To illustrate, 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. 2018). The resulting set of objects (`p_U`, `f_Y`, `I`, `p_S`, `p_Z`, `R`, and `A`) are all steps. Formally each of these steps is a function. The design is the concatenation of these, which we represent using the "+" operator: `design = p_U + f_Y + I + p_S + p_Z + R + A`. A single simulation runs through these steps, calling each of these functions successively. 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

	Design Declaration	Code
M	Declare background variables	`p_U <- declare_population(N = 200, u = rnorm(N))`
M	Declare functional relations	`f_Y <- declare_potential_outcomes(Y ~ Z + u)`
I	Declare inquiry	`Q <- declare_estimand(mean(Y_Z_1 - Y_Z_0))`
D	Declare sampling	`p_S <- declare_sampling(n = 100)`
D	Declare assignment	`p_Z <- declare_assignment(m = 50)`
D	Declare outcome revelation	`R <- declare_reveal(Y, Z)`
A	Declare answer strategy	`A <- declare_estimator(Y ~ Z, estimand = Q)`
	Declare design, <*M*, *I*, *D*, *A*>	`design <- p_U + f_Y + Q + p_S + p_Z + R + A`

	Design Simulation (1 draw)	Code
1	Draw a population u using $P(U)$	`u <- p_U()`
2	Generate potential outcomes using f_Y	`D <- f_Y(u)`
3	Calculate estimand a^M	`a_M <- Q(D)`
4	Draw data, d, given Model assumptions and Data strategies	`d <- R(p_Z(p_S(D)))`
5	Calculate answers, a^A using A and d:	`a_A <- A(d)`
6	Calculate a diagnostic statistic t using a^A and a^M	`t <- a_A["estimate"] - a_M["estimand"]`

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

Table 3: A procedure for declaring and diagnosing research designs using the companion software `DeclareDesign` (Blair et al. 2018). 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.

the range of likely values of the diagnosand compare favorably to reference values such as statistical power of 0.8.[9]

Design diagnosis places 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 often unavoidable when diagnosing designs. In practice, doing so 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

[9]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.

can be mitigated by assessing the sensitivity of diagnosis to alternative models and strategies, which is relatively straightforward given a diagnosand-complete design declaration. Further, researchers can inform their substantive models with already existing data. 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 descriptive, causal, and exploratory research questions can be declared and diagnosed in practice.

3.1 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 of the lack of 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 S2.1, 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 taking a random sample from the national adult population. 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 S2.2, 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 according to this probability. The **I**nquiry pertains 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 0.50 and a standard deviation of 0.29; in the second, the priors place more probability mass at 0.50, with a standard deviation of 0.11. 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%.

3.2 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.

Process Tracing. Although not all approaches to process tracing are readily amenable to design declaration (e.g., theory-building process tracing, see Beach and Pedersen 2013, p. 16), some are. We illustrate focusing on Bayesian frameworks that have been used to describe process tracing logics. In these approaches "causal process observations" (CPOs) are believed to be observed with different probabilities depending on the causal process that has played out in a case. Ideal typical CPOs as described by Van Evera (1997) are "hoop tests" (CPOs that are nearly certain to be seen if the hypothesis is true, but likely either way) "smoking-gun tests" (CPOs that are unlikely to be seen in general but are extremely unlikely if a hypothesis is false) and "doubly-decisive tests" (CPOs that are likely to be seen if and only if a hypothesis is true).[10] Unlike much quantitative inference, such studies often pose "causes-of-effects" inquiries (did the presence of a strong middle class cause a revolution?), and not on "effects-of-causes" questions (what is the average effect of a strong middle class on the probability of a revolution happening?) (Goertz and Mahoney 2012). Such inquiries often imply a hypothesis – "the strong middle class caused the revolution", say – that can be investigated using Bayes' rule in a procedure described by Humphreys and Jacobs (2015) and Fairfield and Charman (2017).

Formalizing this kind of process-tracing exercise leads to non-obvious insights about the

[10] See also Collier, Brady and Seawright (2004), Mahoney (2012), Bennett and Checkel (2014), Fairfield (2013).

tradeoffs involved in committing to one or another CPO strategy ex ante. We declare a design based on a **M**odel of the world in which both the driver, X and the outcome, Y, might be present in a given case either because X caused Y or because Y would have been present regardless of X (or perhaps, some alternative cause was responsible for Y). The **I**nquiry is whether X in fact caused Y in the specific case under analysis (i.e., would Y have been different if X were different?). The **D**ata strategy consists of selecting one case from a population of cases, based on the fact that both X and Y are present, and then collecting two causal process observations. Note that the declaration of the design already illustrates an important point: the case selection strategy informs the answer strategy by enabling the researcher to narrow down the number of causal processes that might be at play. This greatly simplifies the application of Bayes' rule to the case in question.

Importantly, the researcher attaches two different ex ante probabilities to the observation of confirmatory evidence in each CPO, depending on whether X did or did not cause Y. Specifically, the probability of observing the evidence in the first CPO when the hypothesis that X caused Y is true, $Pr(E_1 \mid H)$ is .3, whereas the probability of observing the evidence when the hypothesis is false, $Pr(E_1|\neg H)$ is 0. The first CPO thus constitutes a "smoking-gun" test of H. The second CPO contains evidence that is more likely to be seen when the hypothesis is true, $Pr(E_2 \mid H) = .8$, but observing the second piece of evidence is less informative than observing the first because even when H is false and Y happened irrespective of X, there is some probability of observing the second piece of evidence: $Pr(E_2 \mid \neg H) = .2$. The second CPO thus constitutes a "straw-in-the-wind" test.

Interestingly, diagnosis reveals that a researcher who relied solely on the weaker "straw-in-the-wind" test would make *better* inferences on average than one who relied solely on the "smoking gun" test. One does better relying on the straw because, even if it is less informative when observed, it is much more commonly observed than the smoking gun, which is an informative, but rare, clue. The Collier (2011, 826) assertion that, of the four tests, straws-in-the-wind are "the weakest and place the least demand on the researcher's knowledge and assumptions" might thus be seen as an advantage rather than a disadvantage. In practice of course scholars often seek multiple CPOs, possibly of different strength (see, for example, Fairfield 2013). In such cases, the diagnosis reveals, the learning depends on the ways in which these CPOs are correlated. There

are large gains from seeking two CPOs when they are negatively correlated—for example if they arise from alternative causal processes. But there are weak gains when CPOs arise from the same process. Presentations of process tracing rarely describe correlations between CPO probabilities yet the need to specify these (and the gain from doing so) presents itself immediately when a process tracing design is declared.

Qualitative Comparative Analysis (QCA). One approach to mixed methods research focuses on identifying ways that causes combine to produce outcomes. What, for instance, are the combinations of demography, natural resource abundance, and institutional development that give rise to civil wars? An answer might be of the form: conflicts arise when there is natural resource abundance *and* weak institutional structure *or* when there are deep ethnic divisions. The key idea is that different configurations of conditions can lead to the same outcome (equifinality) and the interest is in assessing which combinations of conditions matter.

Many applications of qualitative comparative analysis use Boolean minimization algorithms to assess which configurations of factors are associated with different outcomes. Critics have highlighted that these algorithms are sensitive to measurement error (Hug 2013). Pointing to such sensitivity, some even go as far as to call for the rejection of QCA as a framework for inquiry (Lucas and Szatrowski 2014).

However, a formal declaration of a QCA design makes clear that these calls unnecessarily conflate QCA answer strategies with their inquiries (see Collier 2014, for a similar argument). Contrary to claims that regression analysis and QCA stem from fundamentally different ontologies (Thiem, Baumgartner and Bol 2016), we show that saturated regression analysis may actually be useful in addressing concerns about measurement error in QCA. This simple proof of concept joins efforts towards unifying QCA with aspects of mainstream statistics (Rohlfing 2018) and other qualitative approaches (Rohlfing and Schneider 2018).

In the Online Appendix we declare a QCA design, focusing on the canonical case of binary variables ("crisp-set QCA"). The **M**odel features an outcome Y that arises in a case if and only if cause A is absent *and* cause B is present ($Y = a * B$)—though we note that the approach extends readily to cases with many causes in complex configurations. For our **I**nquiry we wish to know the true set of configurations of conditions that are sufficient to cause Y. The **D**ata strategy involves measuring and encoding knowledge about Y in a truth table. We allow for some error

in this process. As in Rohlfing (2018), we are agnostic as to how this error arises: it may be that scholarly debate generates epistemic uncertainty about whether Y is truly present or absent in a given case, or that there is measurement error due to sampling variability.

For **A**nswer strategies we compare two QCA minimization approaches. The first employs the classical Quine-McCluskey (QMC) minimization algorithm (see Duşa and Thiem 2015, for a definition) and the second the "Consistency Cubes" (CCubes) algorithm (Duşa 2018) to solve for the set of causal conditions that produces Y. This comparison demonstrates the utility of declaration and diagnosis for researchers using QCA algorithms, who might worry about whether their choice of algorithm will alter their inferences. We show that, at least in simple cases such as this, such concerns are minimal.

We also consider how ordinary least squares minimization performs when targeting a QCA estimand. The righthand side of the regression includes indicators for membership in all feasible configurations of A and B. Configurations that predict the presence of Y with probability greater than .5 are then included in the set of sufficient conditions.

The diagnosis of this design shows that QCA algorithms can be successful at pinpointing exactly the combination of conditions that give rise to outcomes. When there is no error and the sample is large enough to ensure sufficient variation in the data, QMC and CCubes successfully recover the correct configuration 100% of the time. The diagnosis also confirms that saturated regression can recover the data generating process correctly and the configuration of causes estimand can then be computed, correctly, from estimated marginal effects.

This last point is important for thinking through the gains from employing the MIDA framework. The declaration clarifies that QCA is not equivalent to saturated regression: without substantial transformation, regression does not target the QCA estimands (Thiem, Baumgartner and Bol 2016). However, it also clarifies that regression models *can* be integrated into classical QCA inquiries, and do very well. Using regression to perform QCA is equivalent to QMC and CCubes when there is no error, and even slightly outperforms these algorithms (on the diagnosands we consider) in the presence of measurement error. More work is required to understand the conditions under which the different approaches perform comparatively well: for example, it may be that saturated regression fails with sufficiently complex models.

However, the declaration and diagnosis illustrate that there need not be a tension between

regression as an estimation procedure and causal configurations as an estimand. Rather than seeing them as rival research paradigms, scholars interested in QCA estimands can combine the sophisticated machinery developed in the QCA literature to characterize configurations of conditions with the machinery developed in the broader statistical literature to uncover data generating processes. Thus for instance, in answer to critiques that the method does not have a strategy for causal identification (Tanner 2014), one could in principle declare designs in which instrumental variables strategies, say, are used in combination with QCA estimands.

Nested Mixed Methods. A second approach to mixed methods research nests qualitative small N analysis within a strategy that involves movement back and forwards between large N theory testing and small N theory validation and theory generation. Lieberman (2005) describes a strategy of nested analysis of this form. The strategy has many elements of a complete design including strategies for case selection and theory development. In the Online Appendix, we specify the estimands and analysis strategies implied by the procedure proposed in Lieberman (2005). We declare a nested analysis design; we assume a **M**odel with binary variables and **I**nquiry focused on the relationship between X and Y (both causes of effects and effects of causes are studied). The model allows for the possibility that there are variables that are not known to the researcher when conducting large N analysis, but might modify or confound the relationship between X and Y. The **D**ata strategy and **A**nswer strategies are quite complex and integrated with each other. The researcher begins by analyzing a data set in involving X and Y. If the quantitative analysis is "successful" — where in this declaration we define success in terms of sufficient residual variance explained — the researcher engages in within-case "on the regression line" analysis. Using within-case data, the researcher assesses the extent to which X plausibly caused Y (or not X caused not Y) in these cases. If the qualitative or quantitative analyses reject the model then new qualitative analysis is undertaken to better understand the relationship between X and Y. In the design, this qualitative exploration is treated as the possibility of discovering the importance of a third variable that may moderate the effect of X on Y. If an alternative model is successfully developed this is then tested on (the same) large N data again.

Diagnosis of this design illustrates some of its advantages; in particular that in some settings the within-case analysis can guide researchers to models that better capture data generating processes and improve identification. The declaration also highlights the design features that are left

up to researchers. How many cases to gather? What thresholds to use to decide whether a theory is successful or not? The diagnosis of the design we declare suggests interesting interactions between these design elements. For instance, if the bar for success in the theory testing stage is low (in terms of the minimum share of cases explained that are considered adequate) then researcher might be better off sampling fewer qualitative cases in the testing stage and more in the development stage precisely because more variability in the first stage makes it more likely that one would reject a theory when, unknown to the researcher, they would be able to discover a better theory.

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. Sometimes studies describe their goal as the estimation of a parameter β from a model 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 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 that lets us quickly assess the properties of a regression

[11] An alternative might be to imagine an 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.

estimate under the assumption that in the true data-generating process y is in fact a nonlinear function of x (Online Appendix Section S2.6). 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 differing probabilities. The benefit of declaration here is that, without defining I, it is hard to see the conditions under which A is biased or unbiased. Declaration and diagnosis clarify that, even though the answer strategy "assumes" a non-linear relationship in M that does not hold, under certain conditions OLS is still able to estimate a linear transformation of that relationship.

Matching on Observables. In many observational research designs, the processes by which units are assigned to treatment are not known with certainty. In matching designs, the effects of unknown assignment procedure may, for example, be assessed by matching units on their observable traits under an an assumption of as-if random assignment between matched pairs. Diagnosis in such instances can shed light on when such assumptions are justified. In Online Appendix Section S2.7, 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 S2.8, we declare and diagnose such a design. The declaration highlights a difference between this design and many others: the estimand here is not an average of potential outcomes of units, but rather an unobservable quantity defined at the limit of the discontinuity. This feature makes the definition of diagnosands difficult. Assessing bias or external validity is made complicated by the fact that the estimand does not answer a question about actual units. Of course if researchers postulate unobservable counterfactuals (such as the 'treated' outcome for a unit located below the treatment threshold) then the usefulness of the regression discontinuity estimate as an estimate for the average treatment effect for some specific set of units can be assessed.

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 researchers interested in the effect of each of two treatments *conditional on the other treatment being in the control condition*. Should they 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 let the interaction between treatments from $[-.2, .2]$. 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.

We declare and diagnose this design and confirm 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 S2.9. However, as the interaction between the two

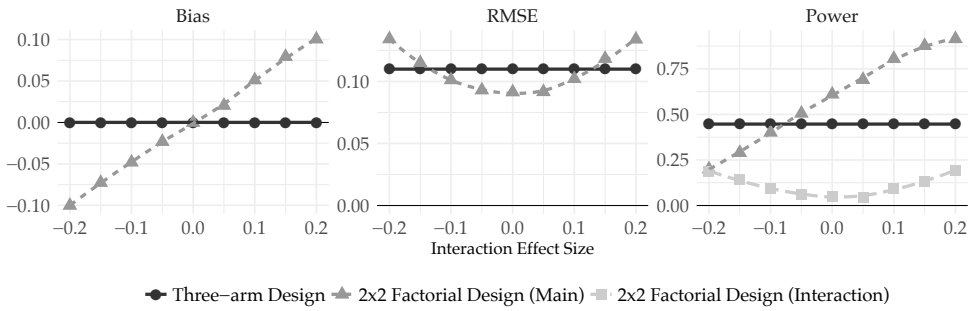

Figure 1: Diagnoses of Designs with Factorial or Three-Arm Assignment Strategies Illustrate a Bias-Variance Tradeoff. Bias (left), root mean-squared-error (center) and power (right) are displayed for two assignment strategies, a 2×2 treatment arm factorial design (black solid lines; circles) and a three-arm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

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 is weak. When the interaction term is 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. An integrated approach to design declaration here illustrates non-trivial interactions between the *D*ata strategy, on the one hand, and the ability of answers (a^A) to approximate the estimand (a^M), on the other.

3.3 Designs for Discovery Research

In some research projects, the ultimate hypotheses that are assessed are not known at 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 discovery process can be described and evaluated.

We examine a data analysis *procedure* in which the researcher assesses 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 S2.10, we declare a design in which the **M**odel stipulates a treatment of interest but also specifies groups for which there might be heterogeneous treatment effects. The main **I**nquiry pertains to the treatment effect but the researcher anticipates that she may be interested in testing for heterogeneous treatment effects if she observes prima facie evidence for it. The **D**ata strategy involves assignment. The **A**nswer strategy involves examination of main effects, but in addition the researcher examines heterogeneous treatment effects inside a random subgroup of the data. If they find evidence of differential effects they specify a new Inquiry which is assessed on the remaining data. The results on heterogeneous effects are compared against a strategy that simply reports discoveries found using complete data, rather than on split data (we call this the unprincipled approach).

We see lower bias from principled discovery than from unprincipled discovery as one might expect. The declaration and diagnosis however also highlight tradeoffs in terms of mean squared error. Mean squared error is not necessarily lower for the principled approach since less data is used in the final test. Moreover the principled strategy is somewhat less likely to produce a result *at all* since it is less likely that a result would be discovered in a subset of the data than in the entire data set. With this design declared, one can assess what an optimal division of units into training and testing data might be given different hypothesized effect sizes.

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 preanalysis 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 design in code 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 preanalysis 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. Reconciliation of designs declared in code can be conducted automatically, by comparing changes to the code itself (a move from the use of a stratified sampling function to simple random sampling) and by comparing key variables in the design such as sample sizes. For diagnosand-complete designs, reconciliation can itself be considered complete (in the sense that all differences have been identified) with respect to that diagnosand.

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

	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.

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 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.

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

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.

5. Application: Design Replication of Björkman and Svensson (2009)

We illustrate the insights that a formalized approach to design declaration can reveal through an application to the design of Björkman and Svensson (2009).

We conduct a "design replication": using available information, we posit a Model, Inquiry, Data and Answer strategy to assess properties of the Björkman and Svensson (2009). This design replication can be contrasted with the kind of "analytic replication" of this study that has been conducted by Donato and Garcia Mosqueira (2016) or the field replication by Raffler, Posner and Parkerson (2018).

The exercise serves three purposes: first, it sheds light on the sorts of insights the design can produce, even without access to the data and code (which, as of writing, were not publicly

available); second, it highlights how difficulties can arise from designs in which the inquiry is not well defined; third, we can assess the properties of replication strategies, notably those pursued by Donato and Garcia Mosqueira (2016) and Raffler, Posner and Parkerson (2018), in order to make clearer the contributions of such efforts.

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.

We briefly outline the steps of our design replication here, and present more detail in the Online Appendix (Section S4).

M We began by positing a model of the world in which unobserved variables, "family health" and "community health," determine both whether infants survive early childhood and whether they are malnourished.

I Our attempt to define the study's inquiry met with a difficulty: the weight of infants in control areas whose lives would have been saved if they had been in the treatment cannot be observed. Unless we are willing to make conjectures about unobservable states of the world (such as the control weight of a child who would not have survived if assigned to the control), we can only define the average difference in individuals' potential outcomes for those children whose survival is unaffected by the treatment: E[Weight(Z=1) - Weight(Z=0) | Alive(Z=0) = Alive(Z=1) = 1]. Of course, we can define our estimand as the difference in average weights for any surviving children in either state of the world: E[Weight(Z=1)|Alive(Z=1) = 1] - E[Weight(Z=0)|Alive(Z=0) = 1]. Note, however, that this estimand can lead to very aberrant conclusions.[13]

[13]Suppose, for example, that only one child with a very healthy weight survived in the control and all children, with weights ranging from healthy to very unhealthy, survived in the treatment. Despite all those lives saved, this estimand would suggest that the treatment has a large negative impact on health.

D As in the original article we stratify sampling on catchment area and cluster-assign households in 25 of the 50 catchment areas to the intervention.

A We estimate mortality at the cluster level and weight-for-age among living children at the household level, as in Björkman and Svensson (2009).

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, given that we are interested in the average difference in potential outcomes defined over observable states of the world.

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 (a^M), whereas the dashed line represents the average answer, i.e. the answer we expect the design to provide given our assumptions ($E[a^A]$). 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 are saved by the treatment. This pulls down average weight outcome in the treatment group.

We draw upon the "robustness to alternative models" criterion (described in the previous section) to argue for an alternative answer strategy that exhibits less bias under plausible conjectures about the world.

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. This approach is equivalent to the "find always-responders" strategy for avoiding post-treatment bias in audit studies (Coppock 2018). 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, if only among this group. 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.

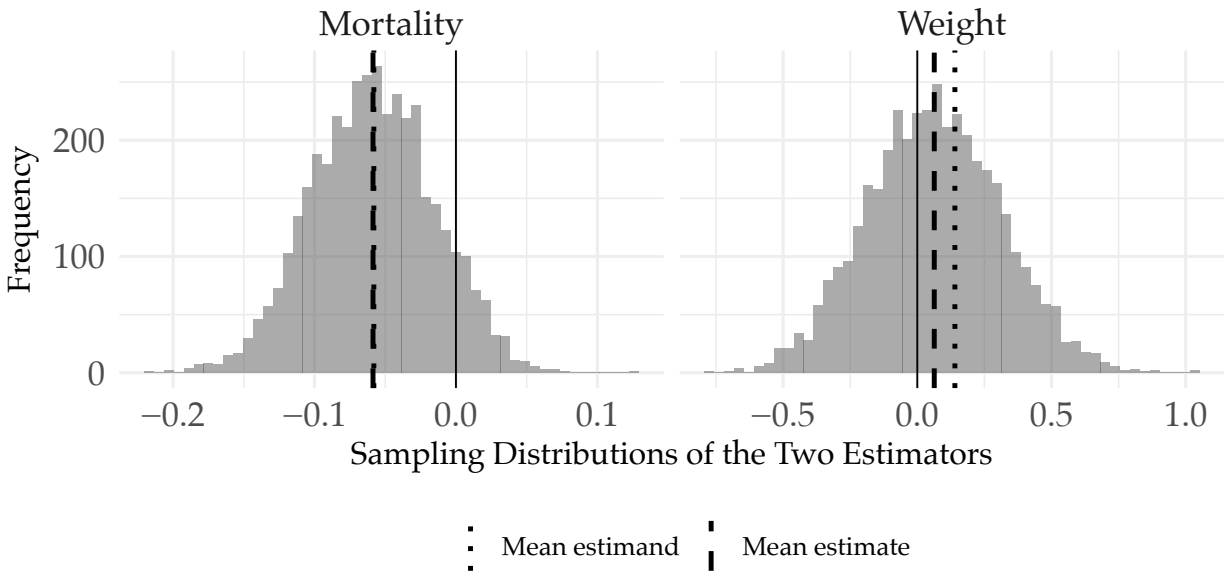

Figure 2: Data-independent replication of estimates in Björkman and Svensson (2009). Histograms display the frequency of simulated estimates of the effect of community monitoring on infant mortality (left) and on weight-for-age (right). The dashed vertical line shows the average estimate, the dotted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

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.

The design replication also highlights the relatively low power of the weight-for-age estimator. As Gelman and Carlin (2014) have shown, conditioning on statistical significance in such contexts can pose risks of exaggerating the true underlying effect size. Based on our assumptions, what can we say here, specifically, about the risk of exaggeration? How effectively does a design such as that used in the replication by Raffler, Posner and Parkerson (2018) mitigate this risk? To answer this question, we modify the sampling strategy of our simulation of the original study to include 187 clusters instead of 50.[14] We then define the diagnosand of interest as the

[14] Raffler, Posner and Parkerson (2018) employ a factorial design which breaks down the original intervention into two subcomponents: interface meetings between the community and village health teams, on the one hand, and integration of report cards into the action plans of health centers, on the other. We augment the sample size here only by the number of clusters corresponding to the pure control and both-arm conditions, as the other conditions of the factorial were not included in the original design. Including those other 189 clusters would only strengthen the conclusions drawn.

"exaggeration ratio" (Gelman and Carlin 2014): the ratio of the absolute value of the estimate to the absolute value of the estimand, given that the estimated effect is significant at the $\alpha = .05$ level. This diagnosand thus provides a measure of how much the design exaggerates effect sizes conditional on statistical significance.

The original design exhibits a high exaggeration ratio, according to the assumptions employed in the simulations: on average, statistically significant estimates tend to exaggerate the true effect of the intervention on mortality by a factor of two and on weight-for-age by a factor of four. In other words, even though the study estimates effects on mortality in an unbiased manner, limiting attention to statistically significant effects provides estimates that are twice as large in absolute value as the true effect size on average. By contrast, using the same sample size as that employed in Raffler, Posner and Parkerson (2018) reduces the exaggeration ratio on the mortality estimand to where it should be, around 1.

Finally, we can also address the analytic replication by Donato and Garcia Mosqueira (2016). The replicators (D&M) noted that the eighteen community-based organizations who carried out the original "power two the people" (P2P) intervention were active in 64 percent of the treatment communities and 48 percent of the control communities. Donato and Garcia Mosqueira (2016) posit that prior presence of these organizations may be correlated with health outcomes, and therefore include in their analytic replication of the mortality and weight-for-age regressions both an indicator for CBO presence and the interaction of the intervention with CBO presence. The inclusion of these terms into the regression reduces the magnitude of the coefficients on the intervention indicator and thereby increases the p-values above the critical $\alpha = 0.1$ threshold in some cases. The original authors (B&S) criticized the replicators' decision to include CBO presence as a regressor, on the grounds that in any such study it is possible to find some unrelated variable whose inclusion will increase standard errors and decrease the coefficient of interest.

In short, the original replicators make a set of contrasting claims about the true *M*odel of the world: B&S claim that CBO presence is unrelated to the outcome of interest, whereas D&M claim that CBO presence might indeed affect (or be otherwise correlated with) health outcomes. As we argued in the previous section, diagnosis of the properties of the answer strategy under these competing claims should determine which answer strategy is best justified.

Since we do not know whether the replicators would have conditioned on CBO presence

and its interaction with the intervention if it had not been imbalanced, we modify the original design to include four different estimation strategies: the first ignores CBO presence as in the original study; the second includes CBO presence irrespective of imbalance; the third includes an indicator for CBO presence only if the CBO presence is significantly imbalanced among the 50 treatment and control clusters at the $\alpha = .05$ level; and the last strategy includes terms for both CBO presence and an interaction of CBO presence with the treatment irrespective of imbalance. We consider how these strategies perform under a model in which, as claimed by the authors, CBO presence is unrelated to health outcomes, and another in which, as claimed by the replicators, CBO presence is highly correlated with health outcomes.

We note first that including the interaction term is a strictly dominated strategy from the standpoint of reducing mean squared error: irrespective of whether CBO presence is correlated with health outcomes or imbalanced, the RMSE expected under this strategy is higher than under any other strategy. Thus, based on a criterion of "Homeground Dominance" in favor of B&S, one would be justified in discounting the importance of the replicators' observation that "including the interaction term leads to a further reduction in magnitude and significance" of the estimated treatment effect (Donato and Garcia Mosqueira 2016, p. 19).

Supposing now that there is no correlation between CBO presence and health outcomes, inclusion of the CBO indicator does increase RMSE ever so slightly in those instances where there is imbalance and the standard errors are ever so slightly larger. On average, however, the strategies of conditioning on CBO presence regardless of balance and conditioning on CBO presence only if imbalanced perform about as well as a strategy of ignoring CBO presence when there is no underlying correlation. However, when there is correlation in health outcomes and CBO presence, strategies that include CBO presence improve RMSE considerably, especially when there is imbalance. Thus, D&M could make a "Robustness to Alternative Models" claim in defense of their strategy: including CBO presence does not greatly diminish inferential quality on average, even if there is no correlation in CBO presence and outcomes; and if there is such a correlation, including CBO presence in the regression specification strictly improves inferences. In sum, a diagnostic approach to replication clarifies that there is very little grounds to update beliefs about the study based on the use of interaction terms, but that the inclusion of the CBO indicator only harms inferences in a very small subset of cases. In general, including it does not worsen

inferences and in many cases can improve them. This approach helps to clarify which points of disagreement are most critical for how the scientific community should interpret and learn from replication efforts.

6. Conclusion

We began with two problems faced by empirical social science researchers: selecting high quality designs and communicating them to others. The preceding sections have demonstrated how the *MIDA* framework can address both challenges. Once designs are declared in *MIDA* terms, diagnosing their properties and improving them becomes straightforward. Because *MIDA* describes a grammar of research designs that applies across a very broad range of empirical research traditions, it enables efficient sharing of designs with others.

Designing high quality research is difficult and comes with many pitfalls, only a subset of which are ameliorated by the *MIDA* framework. Others we fail to address entirely and in some cases, we may even exacerbate them. We outline four concerns.

The first is the worry that evaluative weight could get 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. For instance a simulation-based claim to unbiasedness that incorporates all features of a design is still only good with respect to the conditions of the simulation. Similarly, a power analyses may be useless if implausible parameters are chosen to raise power above some threshold. While MIDA may encourage more faithful-to-theory declarations, there is nothing in the framework that enforces them.

Second, we see a risk that research may get evaluated on the basis of a narrow but perhaps inappropriate set of diagnosands. Statistical power is often invoked as a key design feature – but even well-powered studies that are biased away from their targets of interest are of little theoretical use. The appropriateness of the diagnosand depends on the purposes of the study. As MIDA is silent on the question of a study's purpose, it cannot guide researchers or critics to the appropriate set of diagnosands by which to evaluate a design. An advantage of the approach however is that the choice of diagnosands gets highlighted and new diagnosands can be generated in response to substantive concerns.

Third, emphasis on the statistical properties of a design can obscure the substantive importance of a question being answered or other qualitative features of a design. 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). Our framework can help researchers determine whether a particular design answers a question well (or at all), but it cannot help researchers choose good questions.

Finally, we see a risk that the variation in the suitability of design declaration to different research strategies may be taken as evidence of the relative superiority of different types of research strategies. While we believe that the range of strategies that can be declared and diagnosed is wider than what one might at first think possible, there is no reason to believe 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.

We conclude on a practical note. In the end, we are asking that scholars add a step to their workflow. We want scholars to formally declare and diagnose their research designs both in order to learn about them and to improve them. Much of the work of declaring and diagnosing designs is already part of how social scientists conduct research: grant proposals, IRB protocols, preanalysis plans, and dissertation prospectuses contain design information and justifications for why the design is appropriate for the question. The lack of a common language to describe designs and their properties, however, seriously hampers the utility of those documents for evaluating design quality. We hope that the inclusion of a declaration and diagnosis step to the research process can help address this basic difficulty.

References

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton: Princeton University Press.

Aronow, Peter M. and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.

Balke, Alexander and Judea Pearl. 1994. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*. Elsevier pp. 46–54.

Beach, Derek and Rasmus Brun Pedersen. 2013. *Process-tracing methods: Foundations and guidelines*. Ann Arbor, Michigan: University of Michigan Press.

Bennett, Andrew and Jeffrey T. Checkel, eds. 2014. *Process Tracing*. Cambridge: 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. 2018. "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.

Brady, Henry E. and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, Maryland: Rowman & Littlefield Publishers.

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.

Collier, David. 2011. "Understanding process tracing." *PS: Political Science & Politics* 44(4):823–830.

Collier, David. 2014. "Comment: QCA should set aside the algorithms." *Sociological Methodology* 44(1):122–126.

Collier, David, Henry E. Brady and Jason Seawright. 2004. "Sources of leverage in causal inference: Toward an alternative view of methodology." *Rethinking social inquiry: Diverse tools, shared standards* 2.

Coppock, Alexander. 2018. "Avoiding Post-Treatment Bias in Audit Experiments." *Journal of Experimental Political Science* . Forthcoming.

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

Dawid, A. Philip. 2000. "Causal inference without counterfactuals." *Journal of the American Statistical Association* 95(450):407–424.

Deaton, Angus S. 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Technical report National Bureau of Economic Research.

Donato, Katherine and Adrian Garcia Mosqueira. 2016. "Power to the people? A replication study of a community-based monitoring programme in Uganda." *3ie Replication Papers* 11.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: a Design-based Approach*. Cambridge: Cambridge: Cambridge University Press.

Duşa, Adrian. 2018. *QCA with R. A comprehensive resource*. Springer.

Duşa, Adrian and Alrik Thiem. 2015. "Enhancing the Minimization of Boolean and Multivalue Output Functions With e QMC." *The Journal of Mathematical Sociology* 39(2):92–108.

Fairfield, Tasha. 2013. "Going where the money is: Strategies for taxing economic elites in unequal democracies." *World Development* 47:42–57.

Fairfield, Tasha and Andrew E. Charman. 2017. "Explicit Bayesian analysis for process tracing: guidelines, opportunities, and caveats." *Political Analysis* 25(3):363–380.

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.

Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory building and research design in comparative politics*. Ann Arbor, Michigan: University of Michigan Press.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge: Cambridge University Press.

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.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Goertz, Gary and James Mahoney. 2012. *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton: Princeton: Princeton University Press.

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.

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.

Heckman, James J., Sergio Urzua and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics* 88(3):389–432.

Herron, Michael C. and Kevin M. Quinn. 2016. "A careful look at modern case selection methods." *Sociological Methods & Research* 45(3):458–492.

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

Hug, Simon. 2013. "Qualitative comparative analysis: How inductive use and measurement error lead to problematic inference." *Political Analysis* 21(2):252–265.

Humphreys, Macartan and Alan M. Jacobs. 2015. "Mixing methods: A Bayesian approach." *American Political Science Review* 109(4):653–673.

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. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2):399–423.

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

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

Lieberman, Evan S. 2005. "Nested analysis as a mixed-method strategy for comparative research." *American Political Science Review* 99(3):435–452.

Lohr, Sharon. 2010. *Sampling: design and analysis*. Boston: Brooks Cole.

Lucas, Samuel R and Alisa Szatrowski. 2014. "Qualitative comparative analysis in critical perspective." *Sociological Methodology* 44(1):1–79.

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

Mahoney, James. 2008. "Toward a unified theory of causality." *Comparative Political Studies* 41(4-5):412–436.

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: Cambridge University Press.

Raffler, Pia, Daniel N. Posner and Doug Parkerson. 2018. "The Weakness of Bottom-Up Accountability: Experimental Evidence from the Ugandan Health Sector." Working paper.

Ragin, Charles. 1987. *The Comparative Method. Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

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

Rohlfing, Ingo. 2018. "Power and False Negatives in Qualitative Comparative Analysis: Foundations, Simulation and Estimation for Empirical Studies." *Political Analysis* 26(1):72–89.

Rohlfing, Ingo and Carsten Q Schneider. 2018. "A unifying framework for causal analysis in set-theoretic multimethod research." *Sociological Methods & Research* 47(1):37–63.

Rosenbaum, Paul R. 2002. *Observational Studies*. Springer.

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

Schneider, Carsten Q. and Claudius Wagemann. 2012. *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. Cambridge: Cambridge University Press.

Seawright, Jason and John Gerring. 2008. "Case selection techniques in case study research: A menu of qualitative and quantitative options." *Political Research Quarterly* 61(2):294–308.

Shadish, William, Thomas D. Cook and Donald Thomas Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

Tanner, Sean. 2014. "QCA is of questionable value for policy research." *Policy and Society* 33(3):287–298.

Thiem, Alrik, Michael Baumgartner and Damien Bol. 2016. "Still lost in translation! A correction of three misunderstandings between configurational comparativists and regressional analysts." *Comparative Political Studies* 49(6):742–774.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca: Cornell University Press.

Yamamoto, Teppei. 2012. "Understanding the past: Statistical analysis of causal attribution." *American Journal of Political Science* 56(1):237–256.

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

Appendix

In this appendix, 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, as in an experimental or quasi-experimental design.

$M(1)$ **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.

```
population <- declare_population(N = 1000, u = rnorm(N))
```

$M(2)$ **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.

```
potential_outcomes <- declare_potential_outcomes(Y ~ 0.25 * Z + u)
```

In many cases, the potential outcomes function (or its features) 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, consider a null 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 relative to a true estimand of zero. This approach will not work for diagnosands like power or the Type-S rate. Second, set a series of potential outcomes functions that correspond to competing theories. This approach enables the researcher to judge whether the design yields answers that help adjudicate between the theories.

I **Estimands**. The estimand function τ creates a summary of potential outcomes. In principle, the estimand function can also take realizations of assignments as arguments, in order to calculate post-treatment estimands. Below, the estimand is the Average Treatment Effect, or the average difference between treated and untreated potential outcomes.

```
estimand <- declare_estimand(ATE = mean(Y_Z_1 - Y_Z_0))
```

$D(1)$ **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 `sampling` describes a sampling strategy and not an actual sample.

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

$D(2)$ **The treatment assignment strategy**. 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. We make explicit here an additional step in which the outcome for Y is "revealed" after Z is determined.

```
assignment <- declare_assignment(m = 50)
reveal <- declare_reveal(Y, Z)
```

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.

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

We then declare the design by adding together the elements. Order matters. Since we have defined the estimand before the sampling step, our estimand is the Population Average Treatment Effect, not the Sample Average Treatment Effect. We have also included a `declare_reveal()` step between the assignment and estimation steps that reveals the outcome `Y` on the basis of the potential outcomes and a realized random assignment.

```
design <-
  population + potential_outcomes +
  estimand +
  sampling + assignment + reveal +
  estimator
```

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(estimate - estimand), keep_defaults = FALSE)
```

Diagnosing the design involves simulating the design many times, then calculating the value of the diagnosand from the resulting simulations.

```
diagnosis <- diagnose_design(design = design,
                             diagnosands = diagnosand,
                             sims = 500, bootstrap_sims = FALSE)
```

The diagnosis returns an estimate of the diagnosands, along with other metadata associated with the simulations.

Declaring and Diagnosing Research Designs

Online Appendix

Graeme Blair Jasper Cooper Alexander Coppock Macartan Humphreys

Contents

S1. Declaration of Simple Design from Text (Section 1.2)

The design discussed in Section 1.2 of the text can be formally declared and defined with the following code:

```
b <- 1

f_Y = function(X, Z, b, e_y)  b*X + Z + e_y

model           <- declare_population(N = 1000,
                    e_y = rnorm(N),
                    e_x = rnorm(N),
                    Z   = rnorm(N),
                    X   = Z + e_x,
                    Y   = f_Y(X, Z, b, e_y))

inquiry         <- declare_estimand(
                    effect = mean(f_Y(X, Z, b, e_y) - f_Y(X-1, Z, b, e_y)))

data_strategy   <- declare_sampling(n = 100)

answer_strategy <- declare_estimator(Y~X, model = lm_robust,
                    estimand = "effect")

design <- model + inquiry + data_strategy + answer_strategy

diagnosis <- diagnose_design(design, diagnosands = declare_diagnosands(
                select = c(rmse, bias, mean_estimand)))
```

Estimand Label	Estimator Label	Term	N Sims	RMSE	Bias	Mean Estimand
effect	estimator	X	500	0.51 (0.00)	0.51 (0.00)	1.00 (0.00)

Note that with a design defined it is relatively easy to alter it and assess results. For instance the below alters the design to one with a non linear data generating process. The estimand cannot be read from the dgp directly here but it still calculable.

```
f_Y = function(X, Z, b, e_y)  b*X^2 + e_y
design <- redesign(design, f_Y = f_Y)
```

Note that in this new design there is no confounding but diagnosis would reveal that there is still bias for this estimand resulting from the non linearity of Y in X.

S2. Diagnoses for the Examples in Sections 3.2 and 3.3

The code examples can be downloaded from the internet and run using the free, open source statistical package `R` and the `DeclareDesign` package.

S2.1 Survey Designs

- *M Model*: We posit that voters have a latent probability of voting that is realized as actual voting through a probit process. The probability of voting is positively correlated with support for the Democratic candidate. People reveal their true support honestly, but in general are more likely to turnout to vote than they reveal.
- *I Inquiry*: We wish to know the true support for the Democratic candidate given that the respondent actually votes.
- *D Data Strategy*: We randomly sample 500 respondents.
- *A Answer Strategy*: We estimate the true support for the Democratic candidate among those who will vote by taking the mean of stated support for the Democratic candidate among those who indicate they are likely voters.

S2.1.1 Declaration

```
population <- declare_population(
  N = 1000,
  latent_voting = rnorm(N),
  latent_HRC_support = .1 * latent_voting + rnorm(N) - .1,
  voter = rbinom(N, 1, prob = pnorm(latent_voting)),
  HRC_supporter = rbinom(N, 1, prob = pnorm(latent_HRC_support)),
  likely_voter = rbinom(N, 1, prob = pnorm(latent_voting - 2)))
sampling <- declare_sampling(n = 500)
estimand <- declare_estimand(true_support = mean(HRC_supporter[voter == 1]))
estimator <- declare_estimator(HRC_supporter ~ 1,
                               model = lm_robust,
                               subset = (likely_voter == 1),
                               term = "(Intercept)",
                               estimand = estimand)
descriptive_inference <- population + sampling + estimand + estimator
```

S2.1.2 Diagnosis

```
descriptive_inference_diagnosis <- diagnose_design(
  descriptive_inference = descriptive_inference,
  diagnosands = declare_diagnosands(select = bias),
  sims = sims,
  bootstrap_sims = FALSE)
```

Design Label	Estimand Label	Bias
descriptive_inference	true_support	0.02

S2.2 Bayesian Descriptive Inference

- *M Model*: We posit a population of successes and failures generated through a probit process.
- *I Inquiry*: We wish to know the true probability of success.
- *D Data Strategy*: We sample 10 units.
- *A Answer Strategy*: We estimate empirical priors and a posterior distribution using a beta-binomial model. We compare two estimators, one that uses flat priors, and another that uses priors whose probability mass is centered at .5.

S2.2.1 Declaration

```
n = 10

# M: Model
population <- declare_population(N = 1000,
                                 noise = rnorm(N, -.1, .05),
                                 prob_success = pnorm(noise),
                                 success = rbinom(N, 1, prob_success))

# I: Inquiry
estimand <- declare_estimand(success_prob = mean(prob_success))

# D: Data Strategy
sampling <- declare_sampling(n = n)

# A: Answer Strategy
beta_binom <- function(data,alpha_0,beta_0){n_successes <- sum(data$success)

n_trials <- length(data$success)

alpha <- n_successes + alpha_0 - 1

beta <- n_trials - n_successes + beta_0 - 1

post <- dbeta(seq(0,1,0.005),alpha,beta)

return(data.frame(
  post_mean = alpha / (alpha + beta),
  prior_mean = alpha_0 / (alpha_0 + beta_0),
  post_sd = sqrt((alpha*beta)/(((alpha+beta)^2)*(alpha+beta+1))),
  prior_sd = sqrt((alpha_0*beta_0)/(((alpha_0+beta_0)^2)*(alpha_0+beta_0+1)))))}


estimator_flat_priors <- declare_estimator(handler = tidy_estimator(beta_binom),
                                           alpha_0 = 1,
```

```
                                          beta_0 = 1,
                                          estimand = estimand,
                                          label = "flat priors")
estimator_info_priors <- declare_estimator(handler = tidy_estimator(beta_binom),
                                          alpha_0 = 10,
                                          beta_0 = 10,
                                          estimand = estimand,
                                          label = "informative priors")
# Design
bayesian_design <-
  population + estimand + sampling + estimator_flat_priors + estimator_info_priors
```

S2.2.2 Diagnosis

```
diagnosands <- declare_diagnosands(
  mean_est = mean(post_mean),
  mean_sd = mean(post_sd),
  bias = mean(post_mean - estimand),
  mean_shift = mean(post_mean - prior_mean),
  sd_shift = mean(post_sd - prior_sd),
  keep_defaults = FALSE)

bayesian_estimation_diagnosis <- diagnose_design(
  redesign(bayesian_design, n = c(10,100)),
  diagnosands = diagnosands, bootstrap_sims = FALSE, sims = sims)
```

n	Estimator Label	Mean Est	Mean Sd	Bias	Mean Shift	SD Shift
10	flat priors	0.46	0.14	-0.00	-0.04	-0.15
10	informative priors	0.49	0.09	0.02	-0.01	-0.02
100	flat priors	0.46	0.05	-0.00	-0.04	-0.24
100	informative priors	0.46	0.05	0.00	-0.04	-0.06

S2.3 Process Tracing

Our process-tracing example draws upon the formalizations provided in Humphreys and Jacobs (2015) and Fairfield and Charman (2017).

- *M Model*: We posit a population of 100 cases, each of which does or does not exhibit the presence of some outcome, $Y \in \{0, 1\}$. For the sake of illustration, we will suppose that Y represents the presence or absence of a civil war. Each case also exhibits the presence or absence of some potential cause, $X \in \{0, 1\}$. For example, we might suppose that X represents the presence or absence of natural resources. In our posited model of the world, we specify that $Pr(X = 1) = .7$ for all cases: i.e., some countries do and some countries do not have natural resources. The outcome Y can be realized through four distinct causal relations. First, the presence of X might cause Y, implying that for such cases: if $X = 0$, then $Y = 0$ and if $X = 1$ then $Y = 1$. In other words, civil wars happen in such cases *because* the country has natural resources. Second, the absence of X might cause Y: if $X = 0$ then $Y = 1$ and if $X = 1$ then $Y = 0$ in such cases. In such cases, civil war breaks out *because* the country does not have natural resources, and would not break out if the country had natural resources. Finally, Y might be present irrespective of X or Y might be absent irrespective of X. Continuing our analogy, such countries would have had civil war or peace, irrespective of whether they also had natural resources. We assume that all four causal relations are distributed evenly (with equal probability) throughout the universe of cases specified by our model.

- *I Inquiry*: We wish to know the answer to a "cause of effects" question. Specifically, we wish to know whether a specific case was one in which the presence (absence) of X caused the presence (absence) of Y: did civil war occur in this country because it had natural resources? Denoting the causal hypothesis that the presence of X causes the presence of Y by $H_{X \rightarrow Y}$, we denote our inquiry formally as $Pr(H_{X \rightarrow Y})$.

- *D Data Strategy*: We restrict our attention only to those cases in which both X and Y are present, and select one at random. In other words, we select only those cases in which civil war occurred and natural resources were present. By definition, we know that $Pr(H_{X \rightarrow Y} \mid X \neq Y) = 0$, for example, because $H_{X \rightarrow Y}$ is inconsistent with a data-generating process in which $X = 0, Y = 1$ or $X = 1, Y = 0$. It cannot be that natural resources were the cause of a civil war in a country that had a civil war but no natural resources. What we want to know is whether we see $Y = 1, X = 1$ *because* X caused Y. The inferential challenge is thus to discover whether the reason (though not the only reason) that Y is present is because X caused it to be so. The data strategy is to generate evidence in favor of one or another underlying causal relationship through the use of causal process observations (CPO) tests. In other words, if a country had a civil war because natural resources caused the civil war, there should be observable clues consistent with this hypothesis. The researcher specifies two CPO tests. The first is a ''smoking-gun'': this CPO is believed to arise with probability 0 if X is not the cause of Y, and with probability .3 if X is the cause of Y. The second is a''straw-in-the-wind": if X did not cause Y the researcher still expects to observe this CPO with probability .2, and if X did in fact cause Y the probability of observing the CPO is .8. Thus, whereas the smoking-gun provides rare but definitive proof of the underlying causal relationship, observing the straw-in-the-wind is more likely when the hypothesis that X caused Y is true, but can also happen when this hypothesis is false. For example, if, just prior to the civil war, an armed group was created whose main name, aims and ideology

were centered around the capture and control natural resources, this CPO may constitute a smoking gun. It is extremely unlikely to happen if $H_{X \rightarrow Y}$ is false, but might not happen even if it is true. The national army taking control over natural resources during a civil war is a straw-in-the-wind. This is very likely to happen if the natural resources caused the war, but also somewhat likely even if they did not. Finally, in addition to specifying beliefs about observing the CPOs depending on whether the hypothesis is true or false, the researcher also (implicitly) specifies a belief about the joint probability of observing both CPOs when the hypothesis that X caused Y is true or false. Namely, they specify that the CPOs are independent conditional on the hypothesis being true or false. In terms of our analogy, this is equivalent to assuming that, while it is more likely to observe the argmed group and the national army's takeover of natural resources when resources truly did cause the civil war, this does not imply anything about the probability of observing the national army takeover *given that* the armed group was created. We relax this assumption below and show that it has strong and underexplored implications for process-tracing inferences.

- *A Answer Strategy*: The researcher uses the CPOs in combination with Bayes' rule to update about the probability that X caused Y. In other words, they form a posterior inference, $Pr(H_{X \rightarrow Y} \mid E)$, where E denotes the CPOs they observe. We specify answer strategies for forming this inference. The first simply ignores the CPOs and is equivalent to stating a prior belief without doing any causal process tracing. The second looks only for a straw-in-the-wind, and the third looks only for a smoking-gun. These single-CPO strategies formalize the notion that process-tracing is time-consuming and costly. However, the fourth strategy conditions posterior inferences on both the straw-in-the-wind *and* the smoking-gun, which is consistent with the multiple-CPO strategies of many process-tracing applications (see, for example, Fairfield (2013)).

S2.3.1 Declaration

Here we declare the procedure described above in code.

```
N <- 100
prior_H <- 0.5
p_t1_H <- 0.3
p_t2_H <- 0.8
cor_t1t2_H <- 0
p_t1_not_H <- 0
p_t2_not_H <- 0.2
cor_t1t2_not_H <- 0
Label_1 <- "Smoking Gun"
Label_2 <- "Straw in the Wind"


# M: Model
population <- declare_population(
  N = N,
  causal_process = sample(c('X_causes_Y', 'Y_regardless',
                            'X_causes_not_Y', 'not_Y_regardless'),
                          N,TRUE),
  X = rbinom(N, 1, .7) == 1,
```

```
  Y = (X & causal_process == "X_causes_Y") | # 1. X causes Y
    (!X & causal_process == "X_causes_not_Y") | # 2. Not X causes Y
    (causal_process == "Y_regardless")# 3. Y happens irrespective of X
)

# D: Data Strategy 1
select_case <- declare_sampling(
  strata = paste(X, Y),
  strata_n = c("X0Y0" = 0, "X0Y1" = 0, "X1Y0" = 0, "X1Y1" = 1))

# I: Inquiry
estimand <-
  declare_estimand(did_X_cause_Y = causal_process == 'X_causes_Y')

# D: Data Strategy 2
# Calculate bivariate probabilities given correlation
joint_prob <- function(p1, p2, rho) {
  r <- rho * (p1 * p2 * (1 - p1) * (1 - p2)) ^ .5
  c(
    p00 = (1 - p1) * (1 - p2) + r,
    p01 = p2 * (1 - p1) - r,
    p10 = p1 * (1 - p2) - r,
    p11 = p1 * p2 + r)}
joint_prob_H <- joint_prob(p_t1_H, p_t2_H, cor_t1t2_H)
joint_prob_not_H <- joint_prob(p_t1_not_H, p_t2_not_H, cor_t1t2_not_H)

trace_processes <- declare_step(
  test_results = sample(
    c("00", "01", "10", "11"),1,
    prob = ifelse(rep(causal_process == "X_causes_Y", 4),
                  joint_prob_H,
                  joint_prob_not_H)),
  t1 = test_results == "10" | test_results == "11",
  t2 = test_results == "01" | test_results == "11",
  handler = fabricate)

# A: Answer Strategy
bayes_estimator <- function(data, p_H = prior_H, p_evid_H, p_evid_not_H,
                            label, result) {
    data.frame(
      posterior_H = p_evid_H * p_H /
        (p_evid_H * p_H + p_evid_not_H * (1 - p_H)),
      estimator_label = label,
      estimand_label = "did_X_cause_Y",
      test_results = result
    )}
```

```
no_tests <- declare_estimator(
  handler = bayes_estimator,
  p_evid_H = 1,
  p_evid_not_H = 1,
  label = "No tests (Prior)",
  result = TRUE
)

smoking_gun <- declare_estimator(
  handler = bayes_estimator,
  p_evid_H = ifelse(data$t1, p_t1_H, 1 - p_t1_H),
  p_evid_not_H = ifelse(data$t1, p_t1_not_H, 1 - p_t1_not_H),
  label = Label_1,
  result = data$t1
)

straw_in_wind <- declare_estimator(
  handler = bayes_estimator,
  p_evid_H = ifelse(data$t2, p_t2_H, 1 - p_t2_H),
  p_evid_not_H = ifelse(data$t2, p_t2_not_H, 1 - p_t2_not_H),
  label = Label_2,
  result = data$t2
)

joint_test <- declare_estimator(
  handler = bayes_estimator,
  p_evid_H = joint_prob_H[c("00", "01", "10", "11") %in% data$test_results],
  p_evid_not_H = joint_prob_not_H[c("00", "01", "10", "11") %in% data$test_results],
  label = paste(Label_1, "and", Label_2),
  result = data$test_results
)

# Design
process_tracing_design <-
  population + select_case + trace_processes + estimand +
  no_tests + smoking_gun + straw_in_wind + joint_test

process_tracing_design <- set_diagnosands(
  process_tracing_design,
  diagnosands = declare_diagnosands(
    bias = mean(posterior_H - estimand),
    rmse = sqrt(mean((posterior_H - estimand)^2)),
    mean_estimand = mean(estimand),
    mean_posterior = mean(posterior_H),
    sd_posterior = sd(posterior_H),
    keep_defaults = FALSE
  ))
```

S2.3.2 Diagnosis

First, how do the different inferential strategies perform when we assume that the CPOs arise independently of one another, given the underlying causal process?

```
pt_diagnosis <- diagnose_design(process_tracing_design, sims = sims)
```

Estimator Label	Bias	RMSE	Mean Estimand	Mean Posterior	SD Posterior
No tests (Prior)	0.02	0.50	0.48	0.50	0.00
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Smoking Gun	0.02	0.45	0.48	0.49	0.20
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Straw in the Wind	0.01	0.40	0.48	0.49	0.30
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Smoking Gun and Straw in the Wind	0.01	0.37	0.48	0.49	0.34
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)

Comparing between strategies where a researcher commits ex ante to only search for smoking guns or straws-in-the-wind, the results are somewhat surprising. First, as expected, the standard deviation in posterior inferences generated by the smoking gun strategy is much lower than that provided by the straw-in-the-wind, because the smoking gun provides greater certainty conditional on observing the CPO. However, on average the RMSE is lower when one only searches for straws-in-the-wind, because they are more commonly observed. By this criterion, the straw-in-the-wind strategy actually outperforms the smoking gun strategy.

Clearly, however, RMSE is minimized by conditioning on both CPOs. But here we have specified that the CPOs provide independent information on the true underlying hypothesis: what about when the CPOs are correlated?

We diagnose the design for cases in which the tests are negatively correlated and positively correlated. Negatively correlated CPOs might arise if they result from substitute processes (either from one path or an alternative path). For example, if the national army is less likely to take control of the natural resources precisely when an armed group has declared that it will fight for them. Positively correlated CPOs might arise if they result from common processes (two observations that arise on the same path). For example, if the national army takes control over natural resources precisely because this counters the stated strategic objectives of the armed group.

```
process_tracing_designs <- expand_design(process_tracing_designer,
                                         cor_t1t2_H = c(-.32, +.32))

pt_diagnosis_corr <- diagnose_design(
  process_tracing_designs,
  sims = sims,
  bootstrap_sims = b_sims)
```

cor_t1t2_H	Estimator Label	RMSE
-0.32	Smoking Gun and Straw in the Wind	0.33

cor_t1t2_H	Estimator Label	RMSE
		(0.01)
0.32	Smoking Gun and Straw in the Wind	0.38
		(0.01)

We see that if two CPOs are sought expected errors are lower when these are negatively correlated with each other. This feature arises because the CPO tests carry less independent information when they are positively correlated. To see this, suppose they were perfectly correlated, so that seeing one guaranteed the other would also be present. In this case, there is no additional information gleaned from the observation of one CPO once the other has been observed: they are effectively equivalent tests.

S2.4 Qualitative Comparative Analysis

Our QCA example examines "Crisp Set" QCA in line with the formalizations provided in Ragin (1987) and drawing on the QCA package developed by Thiem and Dusa (2013) (see also Duşa (2018)).

- *M Model*: We suppose there exists N cases. Whether those cases exhibit an outcome, Y, is determined by the configuration of causal conditions that those cases feature. Specifically, the absence of A and the presence of B are necessary and sufficient causes of the presence of Y. We denote this relationship $Y = a * B$, where lower case a indicates the absence of A and upper case B indictates the presence of B. In the code below we use binary indicators for presence and absence, so that $A = 0$ is equivalent to a and $A = 1$ is equivalent to A, for example. The $*$ operator is equivalent to "and". Thus, the causal relationship can be read "Y happens if and only if A is absent and B is present."
- *I Inquiry*: We wish to know the true set of causal configurations that produce Y.
- *D Data Strategy*: We assume that the researcher does not have direct access to Y, but must encode presence or absence in a truth table. We allow for some error in this coding, but make no claim about what this error represents. For example, it may be that scholarly debate generates epistemic uncertainty about whether Y is truly present or absent in a given case, or that there is measurement error due to sampling variability Rohlfing (2018).
- *A Answer Strategy*: We consider two answer strategies initially. The first employs the classical Quine-McCluskey minimization algorithm (see Duşa and Thiem (2015) for a definition) and the second the "Consistency Cubes" algorithm (Duşa (2018)) to solve for the set of causal conditions that produces Y. Further below, we also consider how least squares minimization performs when targeting a QCA estimand. The righthandside of the regression includes indicators for membership in all feasible configurations of A and B. Configurations that yield predictions for Y greater than .5 are then included in the set of sufficient conditions.

S2.4.1 Declaration

We start by declaring the sample size and assume that the outcome is never miscoded.

```
N <- 6
error_rate <- 0

cases <- declare_population(N = N, A = rbinom(N,1,.2), B = rbinom(N,1,.8))

counterfactuals <- declare_potential_outcomes(
  Y ~ 1 * (A == 0 & B == 1),
  conditions = list(A = 0:1, B = 0:1))

estimand <- declare_estimand(true_configuration = "a*B")

true_outcome <- declare_reveal(Y, c(A, B))

code_outcome <- declare_step(
```

```
  mistake = rbinom(N, 1, error_rate) == 1,
  Y_obs = (1 - mistake) * Y + mistake * (1 - Y),
  handler = fabricate)

minimization_algorithm <- function(data, method, label) {
  estimate <- tryCatch(expr = {
    truth_table <- QCA::truthTable(
      data = data,
      outcome = "Y_Obs",
      incl.cut =  .5,
      conditions = c("A","B"))
    solutions <- minimize(truth_table, include = "?", method = method)$solution
    paste(unique(unlist(solutions)), collapse = " + ")
  },
  error = function(e) "NO SOLUTION")
  return(data.frame(estimate = estimate,
                    estimand_label = "true_configuration",
                    estimator_label = label))
}

QMC_estimator <- declare_estimator(handler = minimization_algorithm,
                                   method = "QMC",
                                   label = "Classic Quine-McCluskey")
CC_estimator <- declare_estimator(handler = minimization_algorithm,
                                    method = "CCubes",
                                    label = "Consistency Cubes")

QCA_design <- cases + counterfactuals + estimand + true_outcome +
  code_outcome + QMC_estimator + CC_estimator
```

One draw of the data looks as follows:

ID	A	B	Y_A_0_B_0	Y_A_1_B_0	Y_A_0_B_1	Y_A_1_B_1	Y	mistake	Y_obs
1	0	1	0	0	1	0	1	FALSE	1
2	0	0	0	0	1	0	0	FALSE	0
3	0	1	0	0	1	0	1	FALSE	1
4	0	1	0	0	1	0	1	FALSE	1
5	1	1	0	0	1	0	0	FALSE	0
6	1	1	0	0	1	0	0	FALSE	0

One draw of the estimates looks as follows:

estimate	estimand_label	estimator_label
a*B	true_configuration	Classic Quine-McCluskey
a*B	true_configuration	Consistency Cubes

In principle there is a broad range of interesting diagnosands to explore. These include notions of "coverage" and "inclusion" but also notions of power (Rohlfing 2018). For simplicity we here focus on two. The first is the probability that our answer strategy produces the *exact* set of necessary and sufficient causal combinations that produce the outcome. The second is the probability that our answer strategy fails to provide a an answer at all (no solution to the minimization problem).

```
QCA_diagnosands <- declare_diagnosands(
  correct_rate = mean(estimand == estimate),
  failure_rate = mean(estimate == "NO SOLUTION"),
  keep_defaults = FALSE
)

QCA_design <- set_diagnosands(QCA_design, QCA_diagnosands)
```

S2.4.2 Diagnosis

We diagnose the design.

```
QCA_diagnosis <- diagnose_design(QCA_design, sims = sims)
```

Estimator Label	Correct Rate	Failure Rate
Classic Quine-McCluskey	0.36	0.07
	(0.01)	(0.01)
Consistency Cubes	0.36	0.00
	(0.01)	(0.00)

The first thing to note is that the two algorithms perform relatively similarly. The consistency cubes algorithm does return answers more often than the classic algorithm. However, those answers are not necessarily more accurate. Around 32% of the time, both classic QMC and CCubes return exactly the right causal configuration.

How do these diagnosands change as both the rate of miscoding errors and the sample size change? To investigate this, we can redesign and diagnose our original design.

```
QCA_designs <- redesign(QCA_design, error_rate = c(0,.05), N = c(5,50))

QCA_designs_diagnosis <- diagnose_design(QCA_designs, sims = sims)
```

Estimator Label	error_rate	N	Correct Rate	Failure Rate
Classic Quine-McCluskey	0	5	0.31	0.13
			(0.01)	(0.01)
Consistency Cubes	0	5	0.31	0.01
			(0.01)	(0.00)
Classic Quine-McCluskey	0.05	5	0.27	0.13
			(0.01)	(0.01)
Consistency Cubes	0.05	5	0.27	0.01

Estimator Label	error_rate	N	Correct Rate	Failure Rate
			(0.01)	(0.00)
Classic Quine-McCluskey	0	50	1.00	0.00
			(0.00)	(0.00)
Consistency Cubes	0	50	1.00	0.00
			(0.00)	(0.00)
Classic Quine-McCluskey	0.05	50	0.94	0.00
			(0.00)	(0.00)
Consistency Cubes	0.05	50	0.94	0.00
			(0.00)	(0.00)

Turning first to how the diagnosands vary by N, we see that both algorithms work perfectly when there is no miscoding of the outcome and the sample is large. Measurement error reduces the accuracy of the solutions in quantifiable ways—-that is, a researcher can assess, for any given true data generating process, the probability with which measurement error will yield to a misidentification of causal paths.

How might the use of a simple linear minimization algorithm perform under the existence of miscoding errors? To investigate this, we declare an answer strategy that uses linear regression to find the configuration of causal conditions that produces Y.

```
linear_minimization <- function(data, label) {
  # Get coefficients from saturated model
  betas <- coef(lm(Y_obs ~ A*B, data = data))
  # Assign any dropped coefficients 0
  betas[is.na(betas)] <- 0
  # Linear model prediction of Pr(Y|A,B)
  predictions <- c(
    "a*b" = betas[1],
    "A*b" = betas[1] + betas[2],
    "a*B" = betas[1] + betas[3],
    "A*B" = betas[1] + betas[2] + betas[3] + betas[4])
  # Grab all configurations that predict Pr(Y) > .5
  configurations <- c("a*b", "A*b", "a*B", "A*B")[predictions > .5]
  if(length(configurations) == 0){
    # If no configurations produce the outcome, no solution found
    estimate <- "NO SOLUTION"
  } else {
    # If configurations are found, translate implicants to canonical form
    sum_of_products <- QCA::sop(paste(configurations, collapse = "+"))
    # If configurations are contradictory, no solution found
    estimate <- ifelse(sum_of_products == "", "NO SOLUTION", sum_of_products)
  }
  return(data.frame(estimate = estimate,
                    estimand_label = "true_configuration",
                    estimator_label = label))
  }
```

```
linear_estimator <- declare_estimator(handler = linear_minimization,
                                      label = "Least Squares")
```

We add the estimator to our original design, modify it to include combinations of a larger sample size and error rate as above, and diagnose the designs' properties. Importantly, we declare a design in which $N = 3$, so that the linear regression is estimating more parameters than it has degrees of freedom.

```
QCA_design_with_lm <- QCA_design + linear_estimator

QCA_designs_with_lm <- redesign(QCA_design_with_lm, error_rate = c(0,.20), N = c(3,50))

QCA_with_lm_diagnoses <- diagnose_design(QCA_designs_with_lm,
                                         sims = sims,
                                         diagnosands = QCA_diagnosands)
```

Estimator Label	error_rate	N	Correct Rate	Failure Rate
Classic Quine-McCluskey	0	3	0.10	0.30
			(0.01)	(0.01)
Consistency Cubes	0	3	0.10	0.05
			(0.01)	(0.00)
Least Squares	0	3	0.10	0.30
			(0.01)	(0.01)
Classic Quine-McCluskey	0.2	3	0.05	0.45
			(0.01)	(0.01)
Consistency Cubes	0.2	3	0.05	0.11
			(0.01)	(0.01)
Least Squares	0.2	3	0.05	0.49
			(0.01)	(0.01)
Classic Quine-McCluskey	0	50	1.00	0.00
			(0.00)	(0.00)
Consistency Cubes	0	50	1.00	0.00
			(0.00)	(0.00)
Least Squares	0	50	1.00	0.00
			(0.00)	(0.00)
Classic Quine-McCluskey	0.2	50	0.72	0.00
			(0.01)	(0.00)
Consistency Cubes	0.2	50	0.72	0.00
			(0.01)	(0.00)
Least Squares	0.2	50	0.78	0.00
			(0.01)	(0.00)

Interestingly, the three approaches yield almost exactly the same results when the sample size is small. The main difference resides in the case when the measurement error *and* sample size are large. In such cases, the saturated regression sometimes perform well. The diagnosis suggests that QCA can be fruitfully integrated with regression based estimation.

S2.5 Nested Mixed Methods

We are interested in knowing the answer to an effect of causes question, here the effect of X on Y, though our strategy also examines causes of effects questions in order to help address it. We employ the nested case analysis strategy suggested by Lieberman (2005).

The original description provided by Lieberman is pitched at a high level and so the properties of different tradeoffs are not immediately clear. For example, is it better to invest effort into theory testing or into theory building? What are the consequences of being more or less lenient with regard to the rejection of large- and small-N analyses?

- *M Model*: We posit a model of the world in which X causes Y but the effect is confounded by a variable, W, that the researcher can learn about through qualitative small-N analysis.
- *I Inquiry*: We wish to know the true average effect of X on Y.
- *D Data Strategy* and *A Answer Strategy*: The data strategy and analysis strategy are intertwined. We start with a simple theory, namely that X has similar effects on Y for all units and that the effect of X on Y is unconfounded, and test it on our large-N dataset. The results of this analysis are deemed "satisfactory and robust" if the residual variance from a regression $Y \sim X$ falls below a given threshold, which is set by the researcher as part of the nested strategy. If the the regression is deemed satisfactory, we engage in small-N model-testing by selecting cases on the regression line (i.e., for which $X = 0, Y = 0$ or $X = 1, Y = 1$). We suppose that small-N qualitative analysis reveals whether Y was truly caused by X in the cases studied, but it is time-consuming and costly to do, so it is limited to a small number. If X caused Y in a satisfactory share of cases, which is determined by a threshold set by the researcher as part of their strategy, then the initial model is accepted. If the initial theory fails to explain an adequate share of cases – or if the initial large-N model was deemed to leave too much residual variance unexplained – we move to the theory-building small-N analysis. Here, we suppose that the researcher learns of the variable W, and updates to a new theory in which Y is a function of both X and W if the way that X affects Y is sufficiently different in the presence or absence of W. If there is sufficient evidence that the effect of X on Y is moderated by W, then it is included in a new updated theory. Finally, the researcher returns to the large-N analysis. If they have been led to update their model they use the new one, otherwise they forego answering their research question.

The simplified version of Figure 1 in Lieberman (2005) that we employ here is depicted below.

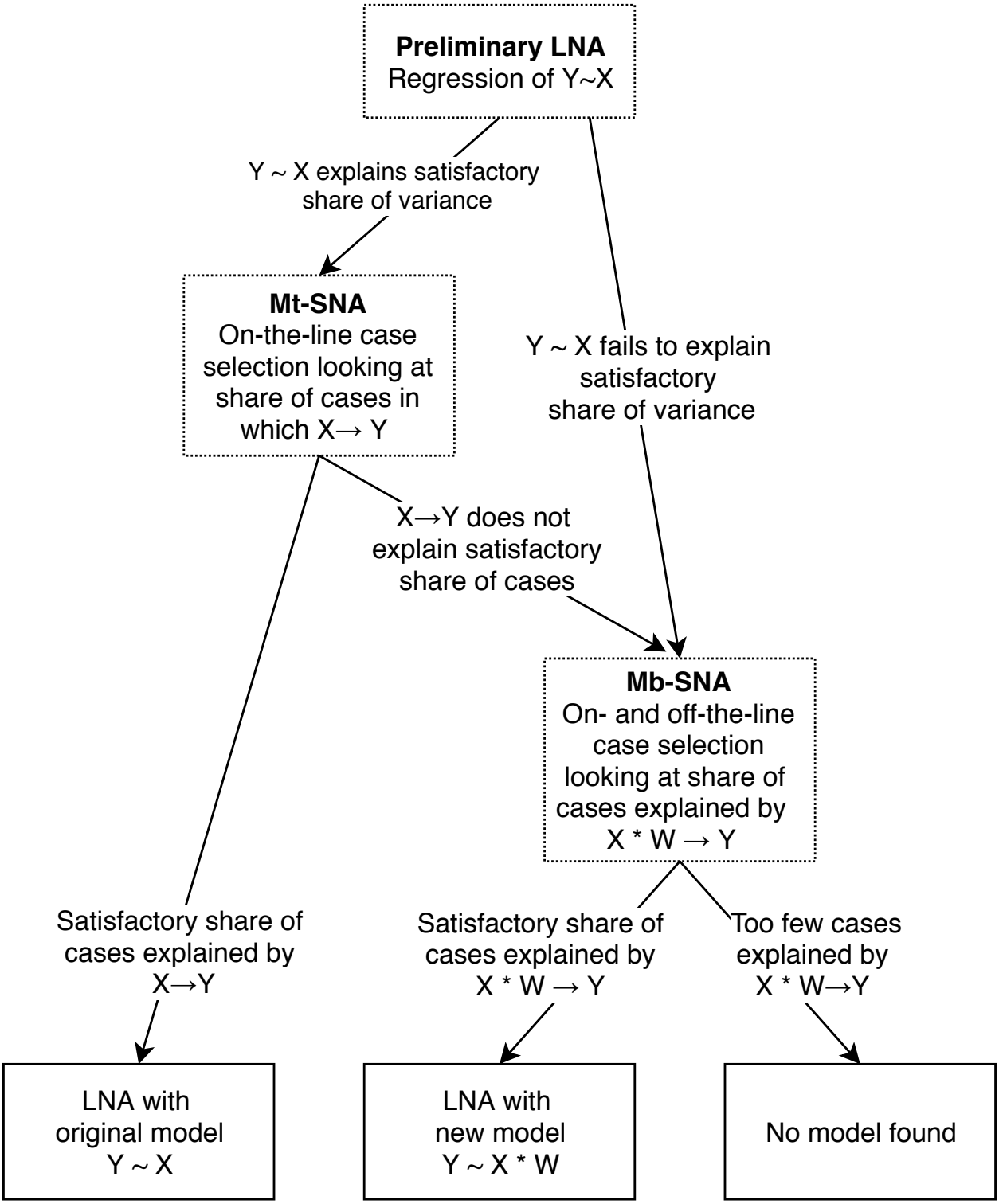

Figure S3: A simplified acyclic version of Lieberman (2005).

S2.5.1 Declaration

```
ate <- 0.5
rho <- 0.7
LNA_threshold <- 0.2
SNA_threshold_1 <- 1/3
SNA_threshold_2 <- 0.1
model_testing_effort <- 1
model_building_effort <- 1

# Model specifies that X causes Y, but confounded by W
LN_data  <-  declare_population(N = 1000,
                                W = rbinom(N, 1, prob = .5),
                                X = correlate(given = W, rho = rho, draw_binary, prob = .5),
                                u = runif(N),
                                Y_X_0 = W*(u > (1 + ate)/2),
                                Y_X_1 = W*(u > (1 - ate)/2),
                                Y = X * Y_X_1 + (1-X) * Y_X_0,
                                outcome = paste0(X,Y))

# We want to know the true average effect of X on Y
true_effect <- declare_estimand(X_on_Y = mean(Y_X_1 - Y_X_0))

# We start with a simple model and see whether X explains enough of Y
LNA <- declare_step(
  LN_res_var = lm_robust(Y~X)$res_var,
  LNA_satisfactory = LN_res_var < LNA_threshold,
  handler = fabricate
)

# Depending on what we find we will select only cases
# "on the Y~X regression line"
select_on_line <- declare_sampling(
  strata = outcome,
  strata_n = model_testing_effort * c("00" = 2, "01" = 0, "10" = 0, "11" = 2))
# ... or "off and on the Y~X regression line"
select_on_and_off_line <- declare_sampling(
  strata = outcome,
  strata_n = model_building_effort * c("00" = 1, "01" = 1, "10" = 1, "11" = 1))

SNA <- declare_step(
  handler = function(data){
    # First, is the LNA deemed satisfactory?
    LNA_satisfactory <- all(data$LNA_satisfactory)
    if(LNA_satisfactory) {
      # If so, select cases on the line
      SN_data <- select_on_line(data)
```

```
    # And determine whether a sufficient share exhibit the posited causal
    # relationship
    SNA_fits_LNA <- mean(with(SN_data,Y_X_1 - Y_X_0)) > SNA_threshold_1
  } else {
    SNA_fits_LNA <- FALSE
  }
  if(LNA_satisfactory & SNA_fits_LNA){
    # If both the LNA and SNA accord with respect to the first model
    # then accept it
    data$situation <- "I: Original Model Accepted"
  } else {
    # Otherwise go back to the full data and sample on and off the line
    SN_data <- select_on_and_off_line(data)
    # Look at effect of X and Y by whether W = 1 or W = 0
    Y_on_X_W <- with(subset(SN_data, W == 1), mean(Y_X_1 - Y_X_0))
    Y_on_X_no_W <- with(subset(SN_data, W == 0), mean(Y_X_1 - Y_X_0))
    # Determine whether the X->Y relationship is moderated by W
    evidence_for_W <- Y_on_X_no_W - Y_on_X_W
    evidence_for_W <- ifelse(is.na(evidence_for_W),0, abs(evidence_for_W))
    new_model_discovered <- evidence_for_W > SNA_threshold_2
    data$situation <- ifelse(test = new_model_discovered,
                             # If W seems to moderate the X->Y
                             # relationship, update the model
                             yes = "II: New Model Accepted",
                             # otherwise, give up
                             no = "IV: Original Model Rejected, No New Model")
  }
  return(data)
})
  # For the final LNA, use the original model if it was kept
  original_LNA <- declare_estimator(
    Y ~ X,
    model = lm_robust,
    estimand = true_effect,
    label = "Original Model")
  # Otherwise use the new model, accounting for X*W interaction
  new_LNA <- declare_estimator(Y ~ X, covariates = ~W,
                               model = lm_lin,
                               estimand = true_effect,
                               label = "New Model")
  final_analysis <- declare_estimator(handler = function(data){
    situation <- unique(data$situation)
    if(situation == "I: Original Model Accepted") return(original_LNA(data))
    if(situation == "II: New Model Accepted") return(new_LNA(data))
    if(situation == "IV: Original Model Rejected, No New Model") {
      return(data.frame(
        estimator_label = "No Model",
```

```
        term = "X",
        estimate = NA,
        std.error = NA,
        statistic = NA,
        p.value = NA,
        conf.low = NA,
        conf.high = NA,
        df = NA,
        outcome = "Y",
        estimand_label = "X_on_Y"
      ))
    }
  })
  # Declare the design
  nested_design <- LN_data + true_effect + LNA + SNA + final_analysis
```

S2.5.2 Diagnosis

We first analyze how confounding poses problems for inference.

```
confounding_designs <- expand_design(nested_designer, rho = c(0,.5))
confounding_diagnosis <- diagnose_design(confounding_designs, sims = sims)
```

rho	Estimator Label	Bias	RMSE	Mean Estimate	Mean Estimand
0	New Model	-0.02	0.02	0.24	0.26
		(0.00)	(0.00)	(0.00)	(0.00)
0	No Model	NaN	NaN	NaN	0.26
		NA	NA	NA	(0.00)
0	Original Model	-0.00	0.02	0.25	0.25
		(0.01)	(0.01)	(0.01)	(0.01)
0.5	New Model	0.01	0.01	0.26	0.25
		(0.00)	(0.00)	(0.01)	(0.01)
0.5	No Model	NaN	NaN	NaN	0.24
		NA	NA	NA	(0.01)
0.5	Original Model	0.15	0.15	0.40	0.26
		(0.01)	(0.01)	(0.01)	(0.00)

As one might expect, the original model is biased if there is truly confounding. Thus, the core question for our nested case strategy is what combination of thresholds and effort devoted to qualitative case analysis will maximize our chances of happening on the correct model?

We look first at how the amount of effort put into different qualitative strategies changes our probability of happening upon the right theory of the $X \rightarrow Y$ relatonship. When `model_testing_effort = 1` and `model_building_effort = 5`, that means we carry out small-N testing in four cases but do theory building in 20, for example.

```
sna_designs <- expand_design(nested_designer,
                             model_testing_effort = c(5,1),
                             model_building_effort = c(1,5))
sna_diagnoses <- diagnose_design(sna_designs, sims = sims)
```

The table below reveals a surprising feature of the nested analysis strategy: increasing the effort devoted to testing *decreases* our chances of happening on the right model. In fact, we are better off when more effort is devoted to qualitative theory building and less effort to qualitative theory testing. The exercise highlights the importance of theory-building.

model_testing_effort	model_building_effort	estimator_label	prop_sims
1	1	New Model	0.3
1	5	New Model	0.5
5	1	New Model	0.2
5	5	New Model	0.3

What can we say about how the weight we put on different forms of evidence affect the likelihood we arrive at the best model? We analyze eight different designs combining low and high thresholds for each of the different decision nodes in our nested strategy. At each, we look at `prop_sims`: the proportion of simulations of the design in which we arive at the correct model.

A low threshold on the LNA means we are less likely to accept an original large-N theory even if it explains a relatively high share of the variance in the outcome. A high threshold for the first SNA test implies we require a large share of cases to exhibit the posited causal relationship in our qualitative model-testing stage in order for us to declare it consistent with the LNA. Finally, a high threshold on the second SNA test implies that we require a large difference in effects in order to believe the alternative theory.

```
threshold_designs <- expand_design(
  nested_designer,
  LNA_threshold = c(.1,.8),
  SNA_threshold_1 = c(.1,.8),
  SNA_threshold_2 = c(.05,.3)
)
threshold_diagnoses <- diagnose_design(threshold_designs, sims = sims)
```

LNA_threshold	SNA_threshold_1	SNA_threshold_2	estimator_label	prop_sims
0.1	0.1	0.3	New Model	0.8
0.1	0.1	0.05	New Model	0.7
0.8	0.8	0.3	New Model	0.7
0.8	0.8	0.05	New Model	0.6
0.1	0.8	0.05	New Model	0.5
0.1	0.8	0.3	New Model	0.4
0.8	0.1	0.05	New Model	0.1

The results suggest that leniency with respect to SNA and LNA testing increase the probabil-

ity of accepting the initial model, and therefore decrease the quality of inferences.

Of course, these insights do not necessarily generalize to alternative nested case analysis strategies, using different thresholds or assuming different inferential problems. However, the declaration shows that it is possible to declare and diagnose very complex inferential procedures whose tradeoffs are otherwise unclear.

S2.6 Observational Regression-Based Strategies

- *M Model*: We posit a population whose potential outcomes are non-linearly but monotonically increasing in some variable, Z.
- *I Inquiry*: We wish to know the average change in potential outcomes brought about by increasing Z by one unit.
- *D Data Strategy*: We imagine two processes through which the values of Z are assigned: in the first, each value is assigned with equal probability; in the second, the highest value is assigned with a lower probability than the lower values.
- *A Answer Strategy*: We estimate the effect of an average unit increase in Z through linear regression of the outcome on Z.

S2.6.1 Declaration

```
# M: Model
population <- declare_population(N = 10, u = rnorm(N))

potential_outcomes <- declare_potential_outcomes(Y_Z_1 = 0 + u,
                                                 Y_Z_2 = 3 + u,
                                                 Y_Z_3 = 4 + u)
# I: Inquiry
model_estimand <- function(Y_Z_1, Y_Z_2, Y_Z_3){
  YY <- c(Y_Z_1, Y_Z_2, Y_Z_3)
  XX <- rep(1:3, each = length(Y_Z_1))
  coef(lm(YY ~ XX))[2]}

estimand <- declare_estimand(
  model_based = model_estimand(Y_Z_1,Y_Z_2,Y_Z_3),
  label = "beta")

# D: Data Strategy
assignment_equal  <- declare_assignment(conditions = 1:3,
                                        prob_each = c(1, 1, 1) / 3)

assignment_unequal  <- declare_assignment(conditions = 1:3,
                                          prob_each = c(.4, .4, .2))

reveal_Y <- declare_reveal()

# A: Answer Strategy
estimator <- declare_estimator(formula = Y ~ Z,
                               model = lm_robust,
                               estimand = estimand,
                               label = "ols")
```

```
# Design
model_estimand_equal <-
  population + potential_outcomes + estimand + assignment_equal +
  reveal_Y + estimator

model_estimand_unequal <-
  population + potential_outcomes + estimand + assignment_unequal +
  reveal_Y + estimator
```

S2.6.2 Diagnosis

We diagnose the two versions of the design with equal and unequal assignment probabilities.

```
model_based_estimand_diagnosis <- diagnose_design(
  model_estimand_equal = model_estimand_equal,
  model_estimand_unequal = model_estimand_unequal,
  bootstrap_sims = b_sims,
  sims = sims)
```

Design Label	Bias	RMSE	Power	Coverage
model_estimand_equal	-0.11	0.25	1.00	1.00
model_estimand_unequal	0.22	0.50	1.00	1.00

The diagnosis reveals that the linear estimator is able to identify a linear estimand, despite both being defined over variables that are non-linearly related to one another. However, as the unequal probability design shows, even with random assignment of Z to Y, the linear estimator can be biased. The diagnosis shows that variation in the assignment of different conditions also matters for inference, in addition to variation in assignment of units.

S2.7 Matching on Observables

- *M Model*: We posit a population that has three standard normally distributed variables, X_1, X_2 and X_3. The potential outcomes of units in the population are an additive function of these variables and the treatment.
- *I Inquiry*: We wish to know the average effect of the treatment among those who were actually treated in a given implementation of the design.
- *D Data Strategy*: We imagine that units are assigned to treatment through a probit process that is a function of the X variables.
- *A Answer Strategy*: We match the units to one another using the three X variables and estimate the difference between treated and control among the matches.

S2.7.1 Declaration

```
population <- declare_population(
  N = 1000, X1 = rnorm(N), X2 = rnorm(N), X3 = rnorm(N))

potential_outcomes <- declare_potential_outcomes(Y ~ X1 + X2 + X3 + Z)

assignment <- declare_assignment(
  handler = function(data) {
    prob <- with(data, pnorm(X1 + X2 + X3))
    data$Z <- rbinom(nrow(data), 1, prob)
    return(data)})

estimand <- declare_estimand(att = mean(Y_Z_1[Z == 1] - Y_Z_0[Z == 1]))

estimator_d_i_m <- declare_estimator(Y ~ Z, estimand = estimand, label = "dim")

match_est <-
  function(data) {
    match_out <- with(data, Match(
      Y = Y,
      Tr = Z,
      X = cbind(X1, X2, X3)
    ))
    return(data.frame(term = "Z", estimate = match_out$est))
  }

estimator_m <- declare_estimator(
  handler = tidy_estimator(match_est),
  estimand = estimand,
  label = "matching")

reveal_Y <- declare_reveal()
```

```
matching <-
  population +
  potential_outcomes +
  assignment +
  reveal_Y +
  estimand +
  estimator_d_i_m +
  estimator_m
```

S2.7.2 Diagnosis

```
matching_diagnosis <- diagnose_design(
  matching, diagnosands = declare_diagnosands(select = bias),
  bootstrap_sims = b_sims, sims = sims)
```

Estimand Label	Estimator Label	Bias
att	dim	2.43
att	matching	0.54

The diagnosis reveals that matching provides a considerable improvement over the naive difference-in-means estimator in terms of bias with respect to the ATT. Nevertheless, even under these ideal conditions the matching estimator fails to provide completely unbiased estimates.

S2.8 Regression Discontinuity

- *M Model*: We posit two potential outcomes functions, one for the treatment condition and another for the control. These functions are fourth order polynomial equations that map the running variable X, to the outcome, Y. We suppose that X is drawn from a uniform distribution, and that units experience an idiosyncratic, normally distributed shock. The treatment variable is 1 when the running variable is greater than .5 (the cutoff) and 0 otherwise.
- *I Inquiry*: We wish to know the true difference in the potential outcomes functions at exactly the point on the running variable where the cutoff is located.
- *D Data Strategy*: We observe the available data without intervening or sampling.
- *A Answer Strategy*: Our estimator is a fourth order polynomial regression in which the terms are fully interacted with the treatment variable.

S2.8.1 Declaration

```
cutoff <- .5
control <- function(X) {
  as.vector(poly(X, 4, raw = T) %*% c(.7, -.8, .5, 1))}
treatment <- function(X) {
  as.vector(poly(X, 4, raw = T) %*% c(0, -1.5, .5, .8)) + .15}
population <- declare_population(
  N = 1000,
  X = runif(N,0,1) - cutoff,
  noise = rnorm(N,0,.1),
  Z = 1 * (X > 0))
potential_outcomes <- declare_potential_outcomes(
  Y_Z_0 = control(X) + noise,
  Y_Z_1 = treatment(X) + noise)
reveal_Y <- declare_reveal()
estimand <- declare_estimand(LATE = treatment(0) - control(0))
estimator <- declare_estimator(
  formula = Y ~ poly(X, 4) * Z,
  model = lm_robust,
  estimand = estimand)
rdd <- population + potential_outcomes + estimand + reveal_Y + estimator
```

S2.8.2 Diagnosis

```
rdd_diagnosis <- diagnose_design(rdd = rdd, bootstrap_sims = FALSE, sims = sims)
```

Estimand Label	Bias	RMSE	Power	Coverage
LATE	16.00	33.14	0.20	0.80

S2.9 Experimental Design

- *M Model*: Our model posits that potential outcomes are a function of two binary treatments, plus, possibly, their interaction combination of background noise and treatment effects. We assume homogeneous effects and in particular, we include individual level shocks but not distinct individual level shocks for each condition.
- *I Inquiry*: We wish to know the effect of treatment 1 compared to the outcome when both treatments are absent.
- *D Data Strategy*: We compare two strategies: a multiarm strategy in which units are assigned to each condition or none (but no units are assigned to have both treatments), and a factorial stragegy in which units are assigned to one of four conditions: no treatment, treatment 1, treatment 2, or both treatments.
- *A Answer Strategy*: We regress the outcome on indicators for both treatment conditions.

In addition we examine the interaction between treatments in the factorial design.

S2.9.1 Declaration

Three-Arm Design

We here use the `multi_arm_designer()` function from the `DesignLibrary` to quickly declare a design that has an N of 500, and three treatment conditions assigned with equal probability, with each non-control treatment generating an effect of .2.

```
multi_arm_design <- multi_arm_designer(N = 500,
                                       m_arms = 3,
                                       outcome_means = c(0, .2,.2))
```

We can quickly inspect the underlying code using `get_design_code(multi_arm_design)`.

2x2 Factorial Design

We similarly use the `simple_factorial_designer()` function to quickly declare a design that has an N of 500, and four treatment conditions assigned with equal probability. Again, each non-control treatment generates an effect of .2 when the other treatment is absent. Their interaction produces an additional interactive effect of .2.

```
simple_factorial_design <-
  simple_factorial_designer(N = 500,
                            outcome_means = c(0,.2,.2,.6))
```

Again, we can quickly inspect the underlying code using `get_design_code(simple_factorial_design)`.

We are interested in comparing the two approaches as the size of the interaction increases. Since we can only estimate the interaction in the 2x2 Factorial, we compare a fixed three-arm design to nine variations of a 2x2 Factorial design.

```
interactions <- seq(-0.2, 0.2, length.out = 9)
means <- lapply(interactions,
                function(interaction) c(0,.2,.2,.4 + interaction))
factorial_designs <- expand_design(
  designer = simple_factorial_designer,
```

```
  prefix = "factorial",
  N = 500,
  w_A = 0,
  w_B = 0,
  outcome_means = means)
```

S2.9.2 Diagnosis

```
factorial_diagnoses <- diagnose_design(factorial_designs,
                                       bootstrap_sims = FALSE,
                                       sims = sims)$diagnosands_df
multi_arm_diagnoses <- diagnose_design(multi_arm_design,
                                       bootstrap_sims = FALSE,
                                       sims = sims)$diagnosands_df
```

interaction	bias_3a	bias_2x2	rmse_3a	rmse_2x2	power_3a	power_2x2_main	power_interaction
-0.20	0.02	-0.12	0.09	0.15	0.5	0.2	0.2
-0.15	0.02	-0.07	0.09	0.09	0.5	0.2	0.1
-0.10	0.02	-0.02	0.09	0.07	0.5	0.6	0.1
-0.05	0.02	-0.04	0.09	0.07	0.5	0.4	0.2
0.00	0.02	-0.02	0.09	0.09	0.5	0.4	0.0
0.05	0.02	0.05	0.09	0.09	0.5	0.8	0.0
0.10	0.02	0.03	0.09	0.09	0.5	0.7	0.0
0.15	0.02	0.08	0.09	0.11	0.5	0.9	0.0
0.20	0.02	0.13	0.09	0.14	0.5	1.0	0.1

The diagnosis reveals confirms that neither design exhibits bias when the true interaction term is equal to zero. 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 (`bias_2x2`). Moreover, there is a bias- variance tradeoff in choosing between the two designs when the interaction is weak. When the interaction term is close to zero, the factorial design is preferred, because it is more powerful (`power_2x2`): it compares one half of the subject pool to the other half, whereas the three arm design only compares a third to a third (`power_3a`). However, as the magnitude of the interaction term increases, the precision gains are offset by the increase in bias.

S2.10 Discovery

- *M Model*: The population consists of two groups (men and women, for instance). The conditional average treatment effect is possibly larger in one group than another.
- *I Inquiry*: The main purpose of the experiment is to estimate the overall Average Treatment Effect. Here, though, we consider the secondary, heterogeneous effects analysis that many researchers conduct after examining the ATE, in which they seek to assess whether effects are larger for one group than for another. This potentially becomes a second question of interest, resulting from discovery.
- *D Data Strategy*: We allocate treatment using complete random assignment.
- *A Answer Strategy*: Using a random half of the data (the test set), we test for an interaction of treatment with group membership. If we find a significant interaction at the $p \leqslant 0.05$ level, we declare the interaction as a new estimand and estimate the size of the interaction in the test data set.

We also include, for comparison, an analysis that reports the interaction estimated using the full data **whenever that interaction is significant**.

S2.10.1 Declaration

```
population <- declare_population(
  N = 200,
  X = draw_binary(prob = rep(0.5, N)),
  het_effect = sample(c(0,.5),1,TRUE),
  train = draw_binary(prob = rep(0.5, N)),
  u = rnorm(N))

potentials <- declare_potential_outcomes(Y ~ Z + het_effect * Z * X + u)
estimand <- declare_estimand(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment()
reveal <- declare_reveal()
main_analysis <- declare_estimator(Y ~ Z, estimand = "ATE", label = "Main")

# Exploration
explore <- declare_step(
  train_pval = coef(summary(lm(Y ~ Z * X, subset = train == 1)))[4,4],
  all_pval = coef(summary(lm(Y ~ Z * X)))[4,4],
  handler = fabricate)

new_estimand <- declare_estimand(
  diff_in_diff = mean(Y_Z_1[X == 1] - Y_Z_0[X == 1]) -
    mean(Y_Z_1[X == 0] - Y_Z_0[X == 0]))

new_estimator <- function(data){
  with(data, data.frame(
    estimate = ifelse(train_pval[1] < .05,
                      coef(lm(Y ~ Z*X, subset = train == 0))[4],
```

```
                      NA),
    p.value = ifelse(train_pval[1] < .05,
                     coef(summary(lm(Y ~ Z * X, subset = train == 0)))[4,4],
                     NA),
    term = "Z:X",
    stringsAsFactors = FALSE))}
new_analysis <- declare_estimator(
  handler = tidy_estimator(new_estimator),
  estimand = new_estimand,
  label = "Discovery")

comparison_estimator <- function(data){
  with(data, data.frame(
    estimate = ifelse(all_pval[1] < .05, coef(lm(Y ~ Z*X))[4], NA),
    p.value = ifelse(all_pval[1] < .05, all_pval[1], NA),
    term = "Z:X",
    stringsAsFactors = FALSE))}
comparison_analysis <- declare_estimator(
  handler = tidy_estimator(comparison_estimator),
  estimand = new_estimand,
  label = "Comparison")

discovery <- population + potentials + estimand + assignment + reveal +
  main_analysis + explore + new_estimand + new_analysis + comparison_analysis

discovery <- set_diagnosands(discovery, declare_diagnosands(
  bias = mean((estimate - estimand), na.rm = TRUE),
  RMSE = sqrt(mean((estimate - estimand)^2, na.rm = TRUE)),
  frequency = mean(!is.na(estimate)),
  false_pos = mean(p.value[estimand == 0] < 0.05, na.rm = TRUE),
  false_neg = 1 - mean(p.value[estimand != 0] < 0.05, na.rm = TRUE),
  keep_defaults = FALSE))
```

S2.10.2 Diagnosis

```
diagnosis <- diagnose_design(discovery, sims = sims*5, bootstrap_sims = b_sims)
```

Estimator Label	Term	Bias	RMSE	Frequency	False Pos	False Neg
Main	Z	-0.00	0.14	1.00	NaN	0.00
Comparison	Z:X	0.38	0.43	0.14	NaN	0.00
Discovery	Z:X	-0.05	0.49	0.14	0.00	0.83

We see that the principled discovery method, using training and testing data, provides essentially unbiased estimates of the heterogeneous effect, whereas the comparison method tends to provide biased estimates, because conditioning on statistical significance tends to exaggerate

effect sizes (Gelman and Carlin (2014)).

The consequences of the two approaches for false discovery rates are stark. If the true effect is 0, the probability of falsely rejecting the null of 0 (conditional on reporting) is 1 under the comparison method: by definition, only significant estimates are kept. Similarly, since the only estimates generated by this procedure are statistically significant, the false negative rate (conditional on reporting) is 0: the estimand is declared and the analysis conducted only when estimates are guaranteed to be significant. The principled discovery method exhibits conventional rates for falsely rejecting a true null (0.05) though it it fails to reject the null quite often, due to weak power.

The protection from bias from the principled discovery strategy declared here does not necessarily translate into improved inferences on average, because of a bias-variance tradeoff inherent in the approach. Less data is used in the final test when adopting a principled discovery approach, and so on average the estimates are much noisier than under the unprincipled comparison.

Moreover the principled strategy is somewhat less likey to produce a result at all since it is less likely that a result would be discovered in a subset of the data than in the entire data set.

With this design, one can assess what an optimal division of units into training and testing data might be given different hypothesized effect sizes.

S3. Further details on survey of design tools

This section describes the construction of the working example used in the research design tool survey, as well as the method used to search for tools to include in the survey, the criteria by which tools were admitted for inclusion into the survey, and the rules for coding the outcomes of this survey. In the online appendix we provide the raw data from the survey, including an overview of the tools considered for inclusion and the reasons for their eventual exclusion, as well as an archive of screenshots of all of the tools included in the survey itself. The tool survey was completed in July 2017 and all findings pertain to the tools we were able to locate at by that time point using the search methods described below.

S3.1 Working Example

There are 1000 city blocks to choose from, each of which contains exactly 25 or 50 households, with the j'th block size distributed categorically, $n_j \sim \text{Cat}(\{25, 50\}, \{.5, .5\})$. Thus, the size of the sample varies as a function of which five city blocks the researcher randomly samples. Specifically, the expected sample size of the study is $N = 5 \times E[n] = 5 \times 37.5 = 187.5$.

Denoting the treatment variable $Z \in \{0, 1\}$, the i'th household respondent's potential outcomes are determined by the following system of equations

$$y_i = Z_i \alpha_j + \epsilon_i, \tag{1}$$

with

$$\alpha_j \sim \text{N}\left(\frac{n_j}{100}, .1\right) \qquad Z_i \sim \text{Bin}\left(\frac{10}{n_j}\right) \qquad \epsilon_i \sim \text{N}(0, 1). \tag{2}$$

Note that the size of the block determines respondents' potential outcomes and their probability of assignment to treatment. Specifically, the two are negatively correlated: the larger the respondent's block, the higher her treated potential outcome and the lower her probability of being assigned to the intervention.

The research design is declared and diagnosed using the following code:

```
set.seed(1:7)
population <- declare_population(
  block = add_level(N = 1000,
                block_size = sample(c(25, 50), N, TRUE),
                block_effect = rnorm(N, block_size / 100, .1)),
  individual = add_level(N = block_size,
                       noise = rnorm(N)))
potential_outcomes <-
  declare_potential_outcomes(formula = Y ~ block_effect * Z + noise)
sampling <- declare_sampling(clusters = block, n = 5)
assignment <- declare_assignment(blocks = block, m = 10)
estimand <- declare_estimand(PATE = mean(Y_Z_1 - Y_Z_0))
dim <- declare_estimator(Y ~ Z,
                         model = lm_robust,
                         label = "DIM",
                         estimand = estimand)
bfe <- declare_estimator(Y ~ Z + block,
                         model = lm_robust,
                         label = "BFE",
                         estimand = estimand)
ipw_bfe <- declare_estimator(Y ~ Z + block,
                             model = lm_robust,
                             label = "IPW-BFE",
                             weights = 1 / Z_cond_prob,
                             estimand = estimand)
reveal_Y <- declare_reveal()
design <-
  population + potential_outcomes + estimand + sampling + assignment +
  reveal_Y +
  dim + bfe + ipw_bfe
```

This code produces the following diagnosis of the design:

```
set.seed(1:7)
diagnosis <- diagnose_design(design, sims = sims, bootstrap_sims = FALSE)
```

Estimator Label	Bias	RMSE	Power	Coverage	Mean Estimate	Mean Estimand
BFE	-0.011	0.179	0.700	0.800	0.406	0.417
DIM	-0.043	0.191	0.600	0.900	0.375	0.417
IPW-BFE	0.008	0.170	0.800	1.000	0.426	0.417

Table S23: Bias, RMSE, power and coverage of design in working example.

Table S23 illustrates that the DIM and BFE estimators are negatively biased: they tend to underestimate the actual size of the treatment effect. This is because it is rarer for units with

high treated potential outcomes to be assigned to treatment, a feature of the design that is not taken into account at all by the DIM estimator, and only through the estimation of a difference in intercepts by the BFE estimator. The IPW-BFE estimator has bias much closer to 0 because it reweights the data to take account of the lower probability with which units in larger blocks are assigned to treatment.

However, the IPW-BFE does not perform strictly better than the BFE estimator in this case. While its power is much higher (72% vs. 58%), this does not result from better efficiency: in fact, the standard deviation of the estimates is higher for the IPW-BFE as a result of the variance introduced by the re-weighting. As the coverage shows, the increased power appears to derive in part from biased variance estimates: the standard errors produced by the IPW-BFE estimator are too small, giving a coverage probability of .88, vs. the more correct coverage probability of the BFE estimator (.93).

In the following sections, we describe the methods by which we sought to assess the ability of available research tools to diagnose these features of the working example design.

S3.2 Search Method

The survey sought to identify computational tools to diagnose the power and bias of the working example design described above. In terms of the identification criteria, we considered any software that promised to design and diagnose prospective research as a candidate for the survey.

We used two principle methods to search for candidates. First, we entered the search terms "statistical bias calculator" , "statistical power calculator" and "sample size calculator" into the Google web search engine, using an incognito browser window in Google Chrome. We assessed the first 30 results using these terms. Second, we assessed the tools listed in four reviews of the literature, namely: Kreidler et al. (2013); Guo et al. (2013); Groemping (2016); Green and MacLeod (2016).

Using these two methods, we identified 143 candidate tools.

Since conducting the survey, the R package PowerUpR was released. We do not include the study here, nor do we include any tool that our search method was able to identify in July 2017. It is thus possible that we missed tools in our original survey. In the case of PowerUpR, while the package could handle power calculations for blocked random assignment as of August 10 2018, it could not incporporate heterogeneous block sizes or assignment probabilities, and thus likely exhibits the same shortcomings described with reference to the other tools surveyed.

S3.3 Admissability Criteria

From the 143 candidate tools, we admitted 30 into the survey. We only admitted those tools that were specifically promised to calculate power or bias in a general purpose way, or in a way that was tailored to the working example. In other words, we excluded tools that were able to calculate power or bias but only for very specific designs that could not accommodate the working example. For instance, the `R` package `ThreeArmedTrials` was a candidate for inclusion because it was listed in the literature review by Groemping (2016) and promised to calculate power of experimental designs. However, because the tool was specifically set up to calculate the power of clinical non-inferiority or superiority trials, we excluded it from consideration in the survey. We also excluded research tools that serve to design research but are not set up to diagnose power or bias. For example, the `experiment` package is set up to design and analyze treatment effects in randomized experiments, but does not provide means for calculating power or bias of designs.

S3.4 Coding Rules

Tools that were included in the survey were coded according to what information on a design they employed to calculate diagnosands (principally bias and power). Some tools accommodated information on design aspects (i.e., block sizes) but did not use this information in the calculation of diagnosands. Tools were only coded as employing a given piece of information if it was included in the calculation of diagnosands.

- *Effect sizes:* When rounded to the third decimal place, the PATE is $\approx .406$ with a standard deviation of 1.01, producing a Cohen's d of approximately .4. Thus, when a tool asked for an effect size without specifying what kind of effect, we entered a value of .4. Sometimes tools require an expression of the effect size in terms of Cohen's f^2. Unlike Cohen's d, the calculation of the f^2 requires that effects be specified in the context of a multivariate regression, and is thus difficult to calculate *a priori*. To calculate the f^2 in this context, we use the companion software to generate 500 R^2 under the full (block FE + treatment) and restricted (block FE only) models, and take the average of the f^2. This is perhaps overly generous to the assessed tools, as the f^2 estimated in this way encodes important design information that the tools do not ask for (such as the assignment probabilities).
- *Heterogeneous block sizes:* 1 if tool allows user to specify that units are organized into groups of different sizes, 0 otherwise.
- *Effect sizes correlated with block sizes:* 1 if tool allows user to specify that effects are correlated with group size, 0 otherwise.
- *Non-constant variance control vs. treatment:* 1 if tool allows for different variances in treatment vs. control, 0 otherwise.
- *Estimand:* 1 if tool allows user to formally define estimand as the Population Average Treatment Effect, 0 otherwise.
- *Sampling strategy:* 1 if tool allows user to specify anything about the strategy via which units are selected from the population into the sample, 0 otherwise.
- *Assign m within blocks:* 1 if tool allows users to specify that exactly m units will be assigned to treatment in the j'th block, 0 otherwise.
- *Inverse-probability weights:* 1 if tool allows users to specify that observations will be weighted by the inverse of their conditional assignment probability during estimation of effects, 0 otherwise.
- *Block fixed-effects:* 1 if tool allows users to specify that a block-level fixed-effect will be estimated, 0 otherwise.
- *Covariate adjustment:* 1 if tool allows users to account for conditioning on covariates, 0 otherwise.
- *Power of DIM:* the estimated power of the difference-in-means estimator if the tool is able to estimate it, NA otherwise.
- *Power of BFE:* the estimated power of the block fixed-effects estimator if the tool is able to estimate it, NA otherwise.

- *Power of IPW-BFE:* the estimated power of the inverse probability-weighted block fixed-effects estimator if the tool is able to estimate it, NA otherwise.
- *Bias:* the estimated bias of any of the estimators if the tool is able to estimate it, NA otherwise.
- *Coverage:* the estimated coverage of any of the estimators if the tool is able to estimate it, NA otherwise.

S4. Bjorkman and Svensson (2009) Design Replication

We present a "design replication" of Björkman and Svensson (2009), by which we mean an exercise in which we learn about the design of a study that has already been conducted. Note that a design replication requires making assumptions about expected features of the data generation processes as well as treatment effects; researchers can disagree on these features. The design replication provides information on features of the design conditional on these assumptions. This exercise is intended to demonstrate how careful specification of estimands can shed light on – and quantify – otherwise hard to assess limitations of analytic strategies.

The study reports the results of a cluster-randomized trial of the effects of community-based monitoring of health clinics in Uganda. The unit of assignment is the health clinic but measurement takes place at the level of the household. Households are considered treated if they are located within the catchment area (5km radius) of a treated health clinic.

The experiment focuses on improvements in two main health outcomes: reductions in child mortality and increases in child weight. The first outcome is measured as the catchment-area-level under-5 mortality rate, expressed in death rates per 1000 live births. In the control group, this rate was 144, compared with 97 in the treatment group: a 33% reduction in child mortality. The second outcome (measured at the household level) is the weight-for-age of infants, defined as children under 18 months. Weight-for-age is measured in standard units, so the positive 0.14 coefficient estimate implies that the weight-for-age of infants in the treatment group was 0.14 standard deviations higher.

We will now characterize this design using the MIDA framework.

S4.1 Model

The population of interest comprises the households within the catchment areas of the 50 health clinics. When we declare the population, we will create three background covariates, two at the household level and one at the catchment area level.

1. `infant`: indicator that equals one if an infant was born into a household in the 18 months preceding the treatment. This variable is observable.
2. `family_health`: a normally distributed variable that represents the health of the household. This variable is likely to be unobservable. We cannot measure it, but it will be positively correlated with the `weight_for_age` of surviving children.
3. `area_health`: a normally distributed variable that represents the overall health of the community. This variable will be the same for all households living within a catchment area and will ensure that outcomes are correlated within catchment area. This variable is also unobservable.

The data are hierarchical – there are 2500 households in each of 50 clusters. The resulting 125,000 row dataset is the population from which subjects will be sampled.

```
# Number of clusters in original study
N_catchment_areas <- 50
# Estimated probability of having a child
infant_prob <- (1135 / (1 - 0.1205)) / 5000

pop <- declare_population(
```

```
  catchment_area = add_level(N = 300,
                             area_health = rnorm(N)),
  households = add_level(N = 2500,
                         infant = rbinom(n = N, 1, prob = infant_prob),
                         family_health = rnorm(N)))

fixed_pop <- declare_population(data = pop())
```

The two outcomes of interest are infant mortality and infant weight. We will first build the infant mortality potential outcomes with a custom function. This custom function builds the probability of an infant surviving in terms of a logistic model, then draws from a binomial distribution using the resulting probabilities. We assume that there is a base rate of survival of approximately 86%, and that treatment increases the probability of survival by approximately 5 percentage points. In logits, this is moving from `plogis(1.81) = 86\%` to `plogis(1.81 + 0.5) = 91\%`. The probability of survival is also positively correlated with the latent health of the household and the health of the community. Finally, if a household does not have an infant, then this potential outcome is undefined. We denote treatment status as $Z = 0$ for control and $Z = 1$ for treatment, hence the condition labels `Z0` and `Z1`.

```
alive_po_function <- function(Z, family_health, area_health, infant) {
  alive <- rbinom(n = length(Z),
                  size = 1,
                  prob = plogis(
                    qlogis(0.86) + 0.5 * Z + family_health + area_health))
  alive[infant == 0] <- NA
  return(alive)}

pos_alive <- declare_potential_outcomes(
  formula =
    infant_alive ~ alive_po_function(Z, family_health, area_health, infant))
```

The second potential outcome is the `weight\_for\_age` of surviving infants. This potential outcome is equal to the latent health of the household for control units. The treated potential outcome is the sum of the latent health and the 0.14 standard deviation treatment effect. Finally, this outcome is masked if the infant dies or if the household does not have an infant.

```
weight_po_function <-
  function(Z, infant_alive_Z_0, infant_alive_Z_1, family_health, area_health){
    weight <- 0.14 * Z + family_health + area_health
    masked <- infant_alive_Z_1 * Z + infant_alive_Z_0 * (1 - Z)
    weight[(masked) == 0 | is.na(masked)] <- NA
    return(weight)}

pos_weight <- declare_potential_outcomes(
  formula =
    weight_for_age ~ weight_po_function(Z, infant_alive_Z_0, infant_alive_Z_1,
                                        family_health, area_health))
```

S4.2 Inquiry

We have two inquiries, the average effect on child mortality (at the cluster level) and the average effect on weight-for-age at the household level.

```
cl_mortality_estimand <- declare_estimand(handler = function(data,label){
  data.frame(
    estimand = with(data,(1 - mean(infant_alive_Z_1,na.rm = T)) -
                      (1 - mean(infant_alive_Z_0,na.rm = T))),
    estimand_label = label,
    stringsAsFactors = FALSE)},
  label = "Mortality")

hh_weight_estimand <- declare_estimand(handler = function(data,label){
  with(subset(data, infant_alive_Z_0 == 1 & infant_alive_Z_1 == 1),
       data.frame(
         estimand = mean(weight_for_age_Z_1 - weight_for_age_Z_0),
       estimand_label = label,
       stringsAsFactors = FALSE))},
  label = "Weight")
```

The second estimand has a complication – it is only defined for a subset of the population. The table below shows four types of infants: Type A (for "Adverse") is alive if in control, but dies if in treatment. Type B ("Beneficial") is just the reverse: the child dies if untreated, but survives if treated. Type C ("Chronic") would die under either condition and Type D ("Destined") would live under either condition. For the first three types, the child dies under one condition, the other or both. This means that the difference in weight potential outcomes is undefined for those types. The difference in weight due to treatment is only defined for Type D infants, those who would survive under either treatment. We therefore define the estimand as being the difference in outcomes for Type D.

This estimand is not recoverable from this design, as we cannot distinguish type A from type D in the control group and type B from type D in the treatment group.

Type	Alive (Z = 0)	Alive (Z = 1)	Weight (Z = 0)	Weight (Z = 1)	Estimand
A	1	0	exists	NA	undefined
B	0	1	NA	exists	undefined
C	0	0	NA	NA	undefined
D	1	1	exists	exists	E[Weight(Z=1) - Weight(Z=0)]

```
hh_weight_estimand <- declare_estimand(handler = function(data, label){
  with(subset(data, infant_alive_Z_0 == 1 & infant_alive_Z_1 == 1),
       data.frame(etimand_label = label,
                  estimand = mean(weight_for_age_Z_1 - weight_for_age_Z_0),
                  stringsAsFactors = FALSE))},
  label = "Weight")
```

S4.3 Data Strategy

Our data strategy includes both the stratified sampling of households by catchment areas and the random assignment of catchment areas to treatment or control. Since the target is 5,000 total households, the study samples 100 households from each catchment area. Assignment to treatment is straightforward: 25 of the 50 clusters receive treatment.

```
cl_sampling <- declare_sampling(clusters = catchment_area,n = N_catchment_areas)
hh_sampling <- declare_sampling(strata = catchment_area, prob = 100/2500)
```

```
assignment <- declare_assignment(clusters = catchment_area)

reveal_outcomes <- declare_reveal(outcome_variables  = c(infant_alive, weight_for_age))
```

S4.4 Answer Strategy

The two estimands require different estimation procedures. For the mortality estimand, we first aggregate the data up to the cluster level, then take the difference in cluster means.

```
aggregate_data <- declare_step(
  handler = function(data){
    aggregate(cbind(infant_alive_Z_0, infant_alive_Z_1, infant_alive, Z) ~ catchment_area,
              FUN  = mean,
              na.rm = TRUE,
              data = data)
  })
cl_mortality_estimator <- declare_estimator(
  model = lm_robust,
  formula = (1 - infant_alive) ~ Z,
  estimand = cl_mortality_estimand,
  label = "Mortality est")
```

The second estimand is at the household level, but we must nevertheless cluster our standard errors by the catchment area. Note that we estimate this quantity among all observed values of `weight\_for\_age`. In the control group, the observed values are a mixed of types A and D, and in the treatment group, the values are a mixture of types B and D. Ideally, we would subset the estimation to include only Type D households, but this information requires knowledge of both the treated and untreated potential outcomes, which is impossible. If potential outcomes are correlated with type (as they are in this simulation), this estimator is biased.

```
hh_weight_estimator <- declare_estimator(weight_for_age ~ Z,
                                         model = lm_robust,
                                         clusters = catchment_area,
                                         estimand = hh_weight_estimand,
                                         label = "Weight est")
```

S4.5 Diagnosis of original design

We now provide the `diagnose_design()` function with the declarations we made above. We will draw a large finite population once, then for each simulation, draw a stratified sample, allocate treatments, reveal outcomes, and conduct the estimation.

```
bjorkman_svensson_design <-
  fixed_pop +
  pos_alive + pos_weight +
  cl_sampling + hh_sampling +
  assignment +
  reveal_outcomes +
  hh_weight_estimator +
  hh_weight_estimand +
```

```
  aggregate_data +
  cl_mortality_estimand +
  cl_mortality_estimator

diagnosis <- diagnose_design(
  design = bjorkman_svensson_design, sims = sims)
```

Estimand Label	Mean Estimand	Mean Estimate	Bias	Power
Mortality	-0.05	-0.05	0.01	0.20
	(0.00)	(0.01)	(0.01)	(0.13)
Weight	0.14	0.05	-0.09	0.10
	(0.00)	(0.07)	(0.07)	(0.09)

The summary of the diagnosis output is presented in the table above. Considering the under-5 mortality rate first, we see that the true population average treatment effect is 0.133 percentage points. In our simulations, we estimate the true standard error to be 0.01, which is close to the standard error reported in the original paper of 0.026. The coverage is correct, at 95%. The simulation presented above shows that we are relatively under-powered for the mortality estimand, only 0.7% of simulations returned a statistically significant result.

Turning next to the weight-for-age analysis, the simulations reveal that our estimator is biased. Because we built into our potential outcomes the assumption that less-health infants were the ones who are most likely to be of type B ("Beneficial"), the treatment group mean is pulled down. Under this assumption, the bias is downwards – our analysis systematically understates the effect on weight-for-age among type D infants, the only type for whom the estimand is defined.

S4.6 Increasing Sample Size

The preceding diagnosis suggests that the original design exhibits low power given the posited model of the data-generating process. Gelman and Carlin (2014) has highlighted how low power can generate bias (even in experiments) if researchers and critics restrict their inferences about the underlying effect to statistically significant estimates. Such bias arises from the so-called "statistical significance filter": only abnormally large effect estimates will be significant when power is low.

To measure the risk of bias arising from consumers applying a "statistical significance filter," we can examine what Gelman and Carlin (2014) refer to as the "exaggeration ratio." The "exaggeration ratio" tells us the expected absolute value of the estimate relative to the absolute value of the estimand, given that the estimate is statistically significant at some level.

We here examine the exggeration ratio, and in particular compare it to the exaggeration ration one might expect from the replication exercise conducted by Raffler, Posner, and Parkerson (2018). Specifically, we analyze the extent of the design improvement that results from the increased number of clusters, given that all of our other assumptions about the original design remain unchanged.

We must first decide how much to augment the sample size. Raffler, Posner, and Parkerson (2018) split the original Power to the People intervention into what they determine to be its two most important components: the provision of information and mobilization of health teams and members of the community, on the one hand, and the implementation of meetings for community members and health staff to plan and raise issues, on the other. They use a factorial design, in

which 95 clusters are assigned to control, 97 to meetings without information, 92 to information without meetings, and 92 to a combination of information and meetings. Thus, while the RPP replication augments the total sample size to 376, compared to 50 clusters in the original, if we compare clusters assigned to directly comparable treatment arms we have the 95 assigned to control and the 92 assigned to the full combination.

We keep the same design as before, but increase the number of clusters to 187.

```
RPP_replication <- redesign(bjorkman_svensson_design, N_catchment_areas = (95 + 92))
```

We diagnose the two designs, declaring the new "exaggeration ratio" diagnosand and setting $\alpha = .1$.

```
filter_diagnosis <- diagnose_design(
  bjorkman_svensson_design,
  RPP_replication,
  diagnosands = declare_diagnosands(
    exaggeration_ratio = mean(abs(estimate[p.value < .1]) / abs(estimand[p.value < .1])),
    select = "power"),
  sims = sims)
```

The diagnoses reveal a substantial improvement but highlight ongoing concerns about statistical significance filters. According to this exercise, the original study risked exaggerating the size of the weight effect by a factor of 4, and the mortality effect by a factor of almost 2. By contrast, increasing the number of clusters as in Raffler, Posner, and Parkerson (2018) improves power and so all but eliminates the risk that statistically significant estimates exaggerate the true underlying effect on mortality. Nevertheless, even with the substantial increase in power, the weight estimate is still at risk of exaggeration, in part due to the bias.

Design Label	N_catchment_areas	Estimator Label	Exaggeration Ratio	Power
bjorkman_svensson_design	50	Mortality est	1.24	0.00
			NA	(0.00)
bjorkman_svensson_design	50	Weight est	NaN	0.00
			NA	(0.00)
RPP_replication	187	Mortality est	1.05	0.90
			(0.09)	(0.09)
RPP_replication	187	Weight est	2.50	0.10
			NA	(0.09)

S4.7 Adding Covariates

In their analytic replication of Björkman and Svensson (2009), Donato and Garcia Mosqueira (2016) (D&M) note that the eighteen community-based organizations (CBOs) who carried out the original "Power to the People" intervention were active in 64 percent of the treatment communities and 48 percent of the control communities. The replicators posit that the presence of CBOs may be correlated with health outcomes, and therefore include in their analytic replication of the mortality and weight-for-age regressions both an indicator for CBO presence and the interaction of the intervention with CBO presence.

The original authors (B&S) criticized the replicators' decision to include CBO presence as a regressor, on the grounds that in any such study it is possible to find some unrelated variable

whose inclusion will increase standard errors or decrease the coefficient of interest.

Expressed in terms of MIDA, we have two conflicting claims about the Model: B&S claim that CBO presence is unrelated to the outcome of interest, whereas D&M claim that CBO presence might indeed affect health outcomes. Moreover, we have the proposal of different answer strategies, with D&M claiming that an indicator for CBO presence and even an interaction of the main treatment indicator with CBO presence should be included in the estimator. How can we assess the grounds for these competing claims?

Since we do not know whether the replicators would have conditioned on CBO presence and its interaction with the intervention if it had not been imbalanced, we modify the original design to include four different estimation strategies: the first ignores CBO presence as in the original study; the second includes CBO presence irrespective of imbalance; the third includes an indicator for CBO presence only if the CBO presence is "significantly" imbalanced among the 50 treatment and control clusters (at the $\alpha = .05$ level); and the last strategy includes terms for both CBO presence and an interaction of CBO presence with the treatment irrespective of imbalance.

```
conditional_estimator <- function(data, strategy, alpha = 0.1){
  imbalanced_fit <- lm_robust(formula = CBO ~ Z,data = data,clusters = catchment_area)
  imbalanced <- imbalanced_fit$p.value["Z"] < alpha
  how_imbalanced <- imbalanced_fit$coefficients["Z"] / imbalanced_fit$coefficients["(Intercept)"
  if(strategy == "ignore") formula <- as.formula(weight_for_age ~ Z)
  if(strategy == "include") formula <- as.formula(weight_for_age ~ Z + CBO)
  if(strategy == "interact") formula <- as.formula(weight_for_age ~ Z + CBO + Z : CBO)
  if(strategy == "include if imbalanced") {
    if(imbalanced) formula <- as.formula(weight_for_age ~ Z + CBO)
    else formula <- as.formula(weight_for_age ~ Z)
  }
  cbind(tidy(lm_robust(formula = formula, data = data, clusters = catchment_area))[2,],
        imbalanced = imbalanced, how_imbalanced = how_imbalanced)
}
ignore_CBO <- declare_estimator(
  handler = tidy_estimator(conditional_estimator),
  strategy = "ignore",
  estimand = hh_weight_estimand,
  label = "Ignore CBO")
include_CBO <- declare_estimator(
  handler = tidy_estimator(conditional_estimator),
  strategy = "include",
  estimand = hh_weight_estimand,
  label = "Include CBO")
interact_CBO <- declare_estimator(
  handler = tidy_estimator(conditional_estimator),
  strategy = "interact",
  estimand = hh_weight_estimand,
  label = "Interact CBO")
include_CBO_if_imbalanced <- declare_estimator(
  handler = tidy_estimator(conditional_estimator),
  strategy = "include if imbalanced",
  estimand = hh_weight_estimand,
  label = "Include CBO if imbalanced")
```

We consider how these strategies perform under a model in which, as claimed by the authors, CBO presence is unrelated to health outcomes, and another in which, as claimed by the replicators, CBO presence is highly correlated with health outcomes.

We firstly specify that the correlation is 0 and add a random draw of CBOs to the catchment area level, and redeclare the design. This is consistent with the claim made by B&S, namely that CBO presence is not correlated with health outcomes.

```
CBO_cor <- 0
fixed_pop_CBO <- declare_population(
  data = pop(),
  catchment_area = modify_level(
    CBO = correlate(draw_handler = draw_binary,
                    given = area_health,
                    rho = CBO_cor,
                    prob = (.64+.48)/2)
  )
)
DM_replication_CBO_ind <-
  fixed_pop_CBO +
  pos_alive +
  pos_weight +
  cl_sampling + hh_sampling +
  assignment +
  reveal_outcomes +
  hh_weight_estimand +
  ignore_CBO +
  include_CBO +
  interact_CBO +
  include_CBO_if_imbalanced
```

We then declare an alternative design in which CBO presence is highly correlated with health outcomes in the catchment area.

```
DM_replication_CBO_cor <- redesign(design = DM_replication_CBO_ind, CBO_cor = .8)
```

And we here assess how well the various answer strategies proposed by the authors fare under the differing assumptions. Importantly, we look both at the overall mean-squared error, and at the mean-squared error when CBO presence is and is not significantly imbalanced at the cluster level. These conditional diagnosands shed light on the consequences of only including a covariate when it is imbalanced.

```
covariate_diagnosis <- diagnose_design(
  DM_replication_CBO_ind,
  DM_replication_CBO_cor,
  diagnosands = declare_diagnosands(
    rmse_balanced = sqrt(mean((estimate[!imbalanced] - estimand[!imbalanced])^2)),
    rmse_imbalanced = sqrt(mean((estimate[imbalanced] - estimand[imbalanced])^2)),
    select = c("rmse")),
  sims = sims)
```

We note first that including the interaction term as done by the replicators is a strictly dominated strategy from the standpoint of reducing mean squared error: irrespective of whether CBO presence is correlated with health outcomes or imbalanced, the RMSE expected under this strategy is higher than under any other strategy (this might not be the case for a strategy that first demeaned CBO presence before interacting so that the coefficient on treatment shoots at

the average effect rather than the effect when CBO = 0). Thus, based on a criterion of "Home-ground Dominance" in favor of B&S, one would be justified in discounting the importance of the replicators' observation that the interaction diminishes the significance of the main effect.

Supposing now that there is no correlation between CBO presence and health outcomes, inclusion of the CBO indicator does increase RMSE ever so slightly in those instances where there is imbalance and the standard errors are ever so slightly larger. On average, however, the strategies of conditioning on CBO presence regardless of balance and conditioning on CBO presence only if imbalanced perform about as well as a strategy of ignoring CBO presence when there is no underlying correlation. However, when there is correlation in health outcomes and CBO presence, strategies that include CBO presence improve RMSE considerably, especially when there is imbalance. Thus, D&M could make a "Robustness to Alternative Models" claim in defense of their strategy: including CBO presence does not greatly diminish inferential quality if you do not agree with their claim about the model, and improves it if you do.

CBO_cor	Estimator Label	RMSE	RMSE Balanced	RMSE Imbalanced
0	Ignore CBO	0.27	0.27	NaN
		(0.04)	(0.04)	NA
0	Include CBO	0.27	0.27	NaN
		(0.05)	(0.05)	NA
0	Include CBO if imbalanced	0.27	0.27	NaN
		(0.04)	(0.04)	NA
0	Interact CBO	0.42	0.42	NaN
		(0.06)	(0.06)	NA
0.8	Ignore CBO	0.35	0.35	0.30
		(0.06)	(0.06)	NA
0.8	Include CBO	0.31	0.32	0.14
		(0.04)	(0.04)	NA
0.8	Include CBO if imbalanced	0.34	0.35	0.14
		(0.06)	(0.06)	NA
0.8	Interact CBO	0.30	0.30	0.26
		(0.05)	(0.06)	NA

References

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): 735–69.

Donato, Katherine, and Adrian Garcia Mosqueira. 2016. "Power to the People? A Replication Study of a Community-Based Monitoring Programme in Uganda." *3ie Replication Papers* 11. 3ie.

Duşa, Adrian. 2018. *QCA with R. a Comprehensive Resource*. Springer.

Duşa, Adrian, and Alrik Thiem. 2015. "Enhancing the Minimization of Boolean and Multivalue Output Functions with E Qmc." *The Journal of Mathematical Sociology* 39 (2). Taylor & Francis: 92–108.

Fairfield, Tasha. 2013. "Going Where the Money Is: Strategies for Taxing Economic Elites in Unequal Democracies." *World Development* 47. Elsevier: 42–57.

Fairfield, Tasha, and Andrew E. Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis* 25 (3). Cambridge: Cambridge University Press: 363–80.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations Assessing Type S

(Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641–51.

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). Wiley Online Library: 493–98.

Groemping, Ulrike. 2016. "Design of Experiments (Doe) & Analysis of Experimental Data."

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. doi:10.1186/1471-2288-13-100.

Humphreys, Macartan, and Alan M. Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109 (4). Cambridge: Cambridge University Press: 653–73.

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). NIH Public Access.

Raffler, Pia, Daniel N. Posner, and Doug Parkerson. 2018. "The Weakness of Bottom-up Accountability: Experimental Evidence from the Ugandan Health Sector."

Ragin, Charles. 1987. *The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Rohlfing, Ingo. 2018. "Power and False Negatives in Qualitative Comparative Analysis: Foundations, Simulation and Estimation for Empirical Studies." *Political Analysis* 26 (1). Cambridge: Cambridge University Press: 72–89.

Thiem, Alrik, and Adrian Dusa. 2013. "QCA: A Package for Qualitative Comparative Analysis." *The R Journal* 5 (1): 87–97.