Three risks in the design and interpretation of conjoint experiments

Macartan Humphreys

2026-01-01

There has been an explosion in the use of survey experiments in political science and related disciplines. But oftentimes there is also concomitant confusion regarding what the goal of a survey experiment is and whether survey experiments can achieve that goal. Focusing especially on conjoint survey experiments, I highlight three interpretative risks. The first relates to the nature of the estimand: whether the experiment aims at a causal estimand, or rather uses causal inference techniques instrumentally to target a descriptive estimand. The goal has implications for the optimal design. The second relates to the use of controls and how inferences hinge critically not just on the distribution of values of controls but on which controls are included, to the point of making all but the most modest causal inferences fraught. The third relates to the license for common claims that survey experiments capture the effects of features of the world, rather than the effects of communications about possible states of the world. I identify conditions that justify this inference from micro to macro estimands, but they are extremely restrictive even when the experimental design closely mimics the features of target elections.

1 Introduction

There has been an explosion in the use of survey experiments in political science. By some accounts, survey experiments are now the most common research design in political science (Torreblanca et al. 2025). Some see survey experiments as providing a method for identifying causal effects that are otherwise hard to identify. As described in one account, researchers are "side-stepping the endogeneity and collinearity concerns that threaten our ability to draw causal inferences using observational data" (Kertzer, Renshon, and Yarhi-Milo 2021).

There is a lot to like about survey experiments, but there is also a risk of confusion about what exactly they contribute. In this paper I highlight three risks of confusion that can arise with some usages of survey experiments.

First, I highlight a confusion regarding the nature of the estimand that survey experiments target. Survey experiments can be used for either **causal inference** (estimating treatment effects) or **descriptive inference** (measuring properties/preferences). But it can sometimes be confusing which of these two worthy goals is the goal in any given experiment. The confusion likely arises in part because very similar designs can be used for either purpose, and in part because survey experimentalists often describe estimands as effects of one kind or another, even if the ultimate inferential target is not an effect. The distinction matters because different goals have implications for how you should set things up and how you should interpret results. For instance, if you are using a survey experiment for descriptive inference, there might be simpler and less noisy strategies available.

Second, I highlight a confusion regarding the role of controls in survey experimentation. For some sorts of survey experiments, such as conjoint experiments or vignette experiments, researchers control many factors at once. The motivation is often described as one of controlling for confounding, enhancing realism, or robustness to contexts. As described by Tomz and Weeks (2013), for instance, varying other features lets them "distinguish the effect of democracy from potential confounders." This may seem curious since randomization, not control, generates independence. Controls play a different function in these experiments, not ensuring unbiased estimates but substantively altering the estimand. This is done in multiple ways. Often researchers do not control just background conditions, but plausibly, also features that are "downstream" to treatments of interest, in a sense clarified below. This can fundamentally alter the interpretation of the estimand, making it inappropriate to pool across experiments using different control sets. The disconnect arises in part because of a confusion between the ideal of randomizing a candidate's attribute and randomizing the signal received by a subject. Although experiments might assign a feature randomly to a profile presented to a subject, this does not mean that respondents know or believe the feature to be exogenous—which might be warranted in settings where attributes themselves are known to be randomized. Rather they might imagine it is informative about other features of a candidate. Thus we might learn that information about corruption, say, affects citizens' beliefs about quality, but not learn that citizens believe that corruption does not affect quality. This feature is not unique to

survey experiments, but can arise in other factorial designs also when one factor is "naturally" downstream from another.

Third, in those cases in which the purpose really is for causal inference, there is a risk of confusion regarding what is being manipulated and so which causal estimand is really being targeted. Ultimately in a survey experiment question wording or survey procedures is being manipulated, not features of the world. From such manipulations we might learn a lot about respondents, but this does not give license to extrapolate to the effects of features of the world except under very stringent conditions. This feature is not unique to survey experiments but arises also, for instance, in audit experiments.

One summary of these concerns is that the promise of survey experiments has been exaggerated. Another though is that the evergreen advice (Lundberg, Johnson, and Stewart 2021) to *know your estimand!* seems especially important when using a survey experiment.

2 Descriptive and causal estimands

Before surveying different types of survey experiment, it is useful to clarify distinctions between causal estimands and descriptive estimands and between measurement and inference.[1]

2.1 Measurement, and causal inference, and descriptive inference

There is a useful distinction often made between measurement and inference. Measurement is about directly observing a quantity that exists in the world; inference is about estimating a quantity that is at least partly unobserved. So you measure your pulse to make inferences about the state of your heart.

You can have causal inference or descriptive inference, so the measurement / inference distinction is not itself about causality. In the same way, *identification* — roughly whether you can nail the quantity of interest if you have enough data — is a problem for inference, but it is not a concern unique to experiments. You can have identification problems for causal estimands or descriptive estimands. And of course it bears repeating: even if a quantity is not identified, you can still learn about it (Tamer 2010).

Recognizing that you are doing inference rather than measurement in turn helps clarify the need for estimates of uncertainty. If you have data from a sample and you are interested in the sample average, and your quantity is measurable, then just measure; no need for standard errors or similar. If you have sample data and you are interested in the *population* average, and your quantity is measurable, then do inference, and also report your standard errors.

A key difference between causal and descriptive estimands is that we generally think that descriptive estimands are, in principle, measurable: they exist, though may be very hard to

[1] See also Ch 14 in Blair, Coppock, and Humphreys (2023) which is structured around this distinction.

measure. Causal estimands however involve counterfactual quantities and cannot be measured, even in principle.

This idea builds on a key idea from the counterfactual model of causation: the causal effect is the difference between two 'potential outcomes' Holland (1986); that is, between two things that do not in fact exist, things that "could have" happened. When we talk about description however we are usually talking about describing properties that we think things *actually have*, like knowledge, beliefs, values, or at a minimum that we are pragmatically committed to treating as if they exist.

Why does this matter? Because it means that if you are interested in causal estimands, then you *have* to do inference. If you are interested in descriptive estimands you may or may not have to do inference. You may be able to measure, but you may have to do inference. That matters because if your interest is description, maybe you can get away without doing inference. Worth checking. Maybe you can ask everyone in your sample if they like coffee and also if they like tea. You don't have to randomly ask half if they like coffee and half if they like tea and infer the values for the full sample based on the 'effect' of the question on the answer! Maybe you will find doing it as an experiment rather than a measurement exercise does not add value; or that the possible gains on some fronts, such as reporting biases, don't make up for the cost of having to do inference.

The distinction between descriptive and causal estimands is not always so sharp though. You might question whether all kinds of properties we might want to describe—preferences, loyalties, and so on—*really* exist and can be described in this sense, even in principle. And you might think that some seemingly describable properties are themselves causal quantities in disguise. For example you might think of preferences as a summary of the effects of options on choices. So you might think of a quantity such as "being a racist" both as a property that someone has and as a summary of how they react as a function of features of people they encounter. These can seem like interchangeable interpretations even if formally you can distinguish between them.

As an analogy, we might think of immunity to a disease as a property that someone has, and want to figure out how many have this immunity. We might even be able to measure features that indicate the property (e.g. sickle cell disease for immunity to malaria). In that case we might want to think of this as a descriptive exercise, and even measure the property in a person. But we might also think of immunity as fundamentally about causal relations: that is, we are really asking about how a person would behave in different conditions. A bit more formally, one might imagine a world in which $X \rightarrow Y \leftarrow A$ (all binary nodes) and functional equation $f : Y = XA$. Then X causes Y if and only if A is present. One way of thinking of the problem is to learn about the causal relations captured by f, the other is to learn about A — the value of a node that captures a property that implies a reaction given a background model. In the case of preferences, the model is likely simple: if individuals rank an option highly they are more likely to select it. Thus, as shown in Figure 1, preferences and features combine to form choices. By altering features and observing choices we learn about preferences.

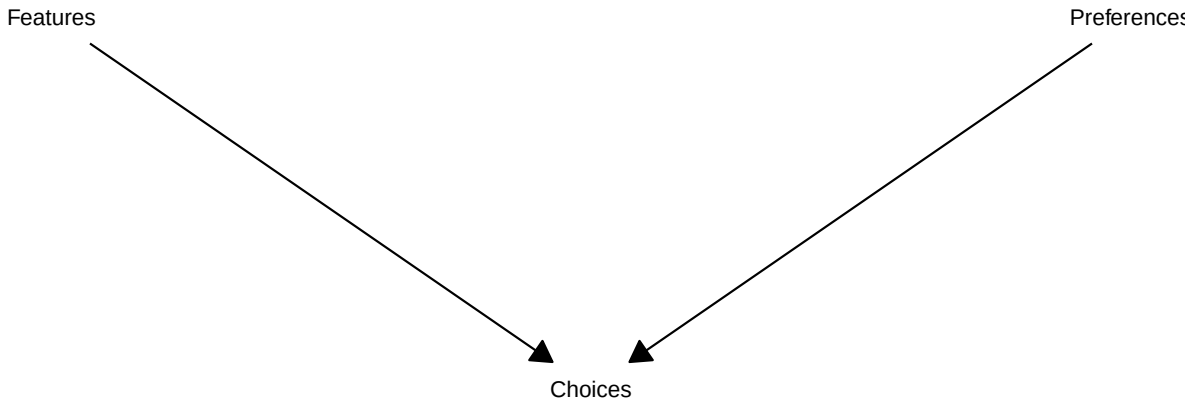

Figure 1: A model in which preferences and features jointly determine choices. Learning about the effect of features on choices lets you make inferences about preferences.

I think three further arguments can often tip towards a property rather than an effect interpretation.

First, in many experiments, choices are not actually made and utility is not actually realized. Rather, individuals make statements about hypotheticals; giving answers that respond to different questions. This is in contrast to audit experiments—which can otherwise look like a conjoint—where a subject may indeed believe an applicant is real and make a decision (in which case the effect of beliefs on behavior is clearly a causal effect). In the survey experiment we record how subjects report they would behave. If we interpret the outcome as how they *would* behave rather than as *what they report* then we assume that respondents know (and truthfully report) potential outcomes, but we do not infer potential outcomes from the realization of the treatment. More modestly we can and perhaps should interpret the "effects" as being on statements, not on utility or actual choices.

Second, in many experiments there is no distinction between the unit and the bundle of treatment conditions: the treatments (features) in a conjoint are *constitutive* of the unit rather than acting upon it. Thus by "changing" a feature you are asking for evaluations on two different units (e.g. evaluation of a corrupt politician versus a clean politician) and not altering a feature of a pre-defined unit (e.g. informing subjects that Jack is in fact corrupt).

Third, the property interpretation might sometimes allow a clearer connection to theory and form a basis for broader inferences. For a related discussion of causal and constitutive *explanations* see Ylikoski (2013). For the key intuition, imagine I asked why your bike is not working. If you because the wheels are missing you are describing the feature of the bike that is preventing a conversion of effort to movement. You are telling me about the "causal capacity" of the bike. If you tell me that it's because turning the pedals does not make the bike go forward you are redescribing the problem. If I am interested in why voters do not vote for female

candidates then one explanation is that voters have preferences against female candidates and vote their preferences; another is that voters do not have preferences against female candidates but believe that others do and their vote would be wasted if they did. Seeing preferences as a property puts you on a path to separate these accounts; seeing preferences as a summary of effects does not.

A weakness of course is the commitment to the idea that the property in question exists. Even a pragmatic commitment to the idea that preferences exist and have explanatory power can itself be reasonably challenged Slovic (1995).

Regardless, which way you think about the estimand can have implications for your design.

Let's now use these ideas to think through different uses of survey experiments.

2.2 Different goals across different types of survey experiment

The term 'survey experiment' is used to cover a large class of experiments. Some are much like any experiment, aiming to estimate a causal effect of a treatment by manipulating that treatment; others use a manipulation, often of survey wording or procedures, to make it possible to measure — or at least make inferences about — a descriptive estimand. See Samii (2025) which makes this distinction, or discussions in Blair, Coppock, and Humphreys (2023). For a counterposition see Schachter and Weisshaar (2025) who suggest that a survey experiment *requires* a causal question.

Sometimes people use the term "survey experiment" specifically for experiments in surveys that use changes in wording or survey protocols to aid descriptive inference, and otherwise say something like "an experiment embedded in a survey" or "delivered through a survey."

In practice though, it is not obvious whether an experiment is conducted to aid descriptive inference or causal inference.

To fix ideas consider two relatively clear cases.

For an example of a survey experiment for **causal inference** consider an **information experiment**. Information experiments are typically used for causal inference, not descriptive inference, whether or not they are delivered through a survey. In some cases survey-delivered information experiments are almost indistinguishable from field experiments — for instance if information is delivered in a way similar to treatments of interest and if outcomes are measured outside of the survey, through measures of subsequent behaviors. The key difficulty with embedding an information experiment in a survey is with respect to external validity—whether the effects of information delivered in this way are similar to effects of information delivered in the wild, and so lots of good work in this vein tries to address that head-on.

For an example of a survey experiment clearly used for **descriptive inference**, consider the **randomized response** survey experiments. In randomized response experiments, people are randomly assigned to answer either a sensitive question or a non-sensitive question and are

typically used for descriptive inference (Blair, Imai, and Zhou 2015). The goal is to estimate the prevalence of some property of subjects, such as whether people have engaged in illegal behavior. The randomization makes it possible to make inferences about the prevalence of the sensitive behavior while protecting individual privacy. Here the randomization is a tool to make measurement possible, not the focus of interest itself. There is a causal effect of the procedure on the answer, but the purpose is to make descriptive inferences about something else.

I think these two cases show a sharp difference between the two goals. The purpose is not always so clear, however. Table 1 summarizes distinct uses to which different types of survey experiment might be put, highlighting traps to avoid if attracted by this type of survey experiment.

Table 1: Summary of different uses for survey experiments

Survey Experiment Type	Causal Inference Use Case	Descriptive Inference Use Case
Priming experiments	Estimate effect of prime on behavior/attitudes (typical)	Use prime as diagnostic to infer knowledge/beliefs (rarer)
List experiments	Estimate effect of list length or content on response patterns (rare)	Infer prevalence of sensitive beliefs/behaviors (typical)
Framing experiments	Estimate effect of politician framing on voter choices (common)	Infer underlying preference purged of framing effects (common)
Conjoints	Estimate effect of feature on choices, given a distribution of other fixed features (rare?)	Make inferences about preferences, classification rules, or ideal points (typical?)

I use question marks in the last row because I am confused on what some of these are trying to do (see examples below).

The next section unpacks the ideas in this table for the case of conjoint experiments. In the Appendices I discuss other types of survey experiments.

2.3 Conjoints

De la Cuesta, Egami, and Imai (2022) describe conjoints as "a factorial survey experiment that is designed to measure multidimensional preferences". Note the emphasis on measurement. In a similar way, Bansak et al. (2023) describes the (AMCE) estimand as a "*summary* of voters' multidimensional preferences" (emphasis added). Arguably, the remit of conjoints for descriptive inference is a little broader. For example they might also be used to study how people make classifications or understand concepts. But, arguably, conjoints might sometimes also be used when the estimand really is causal. The dual usage is rarely recognized however — as highlighted elegantly for instance by Ganter (2023).

2.3.1 Conjoints for descriptive inference.

The conjoint design is a very powerful tool for learning about preferences, interpretations, or classification rules, letting you learn about complex attribute spaces using unobtrusive questions. When used for this purpose, conjoint experiments may be best thought of as using causal inference to make descriptive inferences. See the discussion of the conjoint design in Blair, Coppock, and Humphreys (2023).

For example, in Hartmann et al. (2024), we use a conjoint because we want to measure policy preferences, under different contingencies. We combine the conjoint results with a choice model to estimate ideal points. Although we use the language of effects a bunch we are interested in trying to measure something, but given the complexity of the space and the practical inability to explore it all, resort to using the conjoint to make inferences.

Another example, constructed to highlight the descriptive nature of the exercise: say a bank uses a rule to decide whether to give loans or not. You want to figure out the rule. So you use a conjoint to assess which profiles are more likely to get loans given different attributes. The estimand of interest is not a set of causal effects, it is a rule. But you try to figure it out by seeing whether notional features "affect" the classification. By analogy when you observe stated preferences for different profiles you can use these to figure out the underlying function—rule—that evaluates the profiles, not trying to figure out preferences over the profiles themselves).

Two implications from recognizing that the goal here is in fact descriptive inference:

- Opportunity. You might find out that a more effective strategy would be to figure out the rule from archival sources, such as regulations or instructions to staff. Maybe it is measurable, in which case measure it.
- Risk. You might fall into the trap of thinking the relation between feature values and outcomes corresponds to the causal effects of changing the feature (or confuse the direct/controlled effect within the experimental regime with the average effect). This is a little trickier, but to think through a simple example: Say in truth we have $A_1 \rightarrow A_2 \rightarrow Y$, and A_1 affects Y via A_2 but not conditional on A_2. Then a conjoint might pick up that A_1 is not part of the classification rule for Y and A_2 is. But it would be wrong to infer from this that actually changing A_1 will not affect classifications (since it might via changes in A_2). The problem here is confusing "how the rule determines outcomes given features" with "the effect of changing features, given the rule."

I think when Schwarz and Coppock (2022) talk about learning about discrimination, they are focused on uncovering preferences in this way; but the language of describing "the average effect of *being* a woman" (emphasis added), could be misread to suggest an interest in the effect of the attribute itself, that is, an effect of an intervention on a candidate.

2.3.2 Conjoints for causal inference

Even still, conjoints can also be used when the primary target is a causal estimand. Say you really are interested in whether the presence of a given feature on a list of features makes it more likely that an outcome will be selected from the list.

You might have an application where people are electing candidates and know nothing about the candidates other than what they get in a flyer. You might have a pool of potential candidates with a set of features from which you would draw a pair of potential candidates. You want to know how the presence of a given feature on the flyer affects the choice, conditional on all other features, averaged. Although your interest is the effect of the signals on choices, not the underlying preferences, you are pretty close to the conjoint. You have to worry about external validity but these are common worries for any experiment.

Note that one bonus of your interest being in the "choice" rather than preferences, is that you might not be concerned if you found that people didn't take the exercise too seriously, or didn't read options carefully, as that is just a part of what creates the mapping from features to choices and may be true in the real world also.

I think this is close to the sort of setting Bansak et al. (2023) have in mind (though, note this means interpreting their language of "the effect of a change in an attribute on a candidate's or party's expected vote share" as meaning – as they clarify elsewhere — the effect of a listed feature within a controlled list of features and not the effect of an intervention on a single feature of a candidate while allowing (or preventing) other endogenous changes). And this is more or less the setting examined by Hainmueller, Hangartner, and Yamamoto (2015) where there is a striking parallelism between the application and the survey experimental setting.[2]

The risk above remains, however: the effect you are getting is the effect of the attribute signal on the list, not the average (total) effect of the attribute itself on the outcomes. For example you might find that a powerful candidate does well *given* different values of corruption (even for different distributions of corruption), but this does not give you the effect of power itself, since, after all, power corrupts. You might of course really be interested in effects like this: the causal effect of the candidate's attribute itself, in the sense of imagining the effect of an intervention on a candidate (e.g. the effect of the candidate's wealth on their performance), rather than on the listing of a particular attribute value in a list of attributes. That's a quantity that the conjoint would struggle to identify (see below).

[2]What is striking about the application is its atypicality: that it is a real world setting that is similar to a conjoint—with many candidates but a constrained information set available to voters. Worth noting in this case that the argument for a natural experimental benchmark is not that attributes of candidates are in anyway randomly assigned but that researchers have access to the same information as voters. This feature lets one take account of other attributes in a similar way to what is done in a conjoint, though it does not, in itself, provide identification for the effects of listed attributes. For instance one might imagine that more educated candidates figured out how to come up for a vote at times where voters were more generally favorable to immigrants and avoid times when they were less favorable. This could produce a non causal correlation between the education attribute and voter support that is not addressed by conditioning on other features of the information set.

[[In addition Ganter (2023) argues that with a descriptive goal in mind the researchers should not focus on the AMCE, which, he argues, is more suited to selection-process estimands. The intuition—from Ganter (2023)—can be communicated with the example in which voters have a preference for women over men. This preference is fixed. However the advantage of being female disappears if in selection choices women are in fact paired with women and men with men. Thus gender may be irrelevant to the choice, given the options; be prominent in preferences. I address this point from a different angle later, suggesting that "preferences" in these contexts are often best conceived as menu dependent.]]

3 The role of controls

Conjoint experiments are celebrated not least because of the ability to add many controls in a simple and natural way – both the ability to control multiple items in the delivery of the treatment and the ability to include controls in the analysis. But it can sometimes be unclear what role controls play.

3.1 What roles do the controls play?

Controlling features can mean different things in experimental research. To clarify terminology, we can think of controlling at the intervention stage (fixing experimental conditions) or at the analysis stage (taking account of third features when we estimate effects). The former is standard in lab experimental settings, but it also arises in field experimental settings when researchers use factorial designs.[3] Controls in the analysis stage are used for three distinct purposes. To remove bias or confounding in the estimation of average effects, to improve precision (reduce variance); and to target specific quantities of interest, such as conditional or controlled effects. In experiments it is well known that controls are not needed for the first purpose, and can indeed introduce bias. But controls are regularly used for the latter two purposes.

The bias argument. Interestingly, controls, in survey experiments are sometimes described as helping to address confounding. This is a somewhat curious claim since random assignment itself ought to remove confounding—a key merit of random assignment is that there is reduced need to introduce controls. Random assignment should remove bias, at least if the treatment of interest is the treatment assigned. But researchers worry about confounding all the same.

[3]Fisher (1971) made three arguments for *varying* conditions at intervention stage [p106]: 1. Efficiency—in that multiple treatments can be assessed with the same observations 2. Comprehensiveness—additional estimands, such as interaction effects, can be estimated. 3. Widening of scope—statements can be made about effects that are not a function of specific standardization decisions. He highlighted that standardization "weakens rather than strengthens our ground for inferring a like result" and so variation allows for a wider inductive basis. The idea taken up more recently is that such variation allows one more easily to make out-of-sample predictions to different contexts. See De la Cuesta, Egami, and Imai (2022) and Tipton (2021) for implications of a related argument for sampling.

Tomz and Weeks (2013) describe a concern about leaving out factors "that could confound the relationship between shared democracy and public support for war." Bell and Quek (2018) are a little more specific and write that by "holding the military power of the target constant, we reduce the possibility of the respondents drawing inferences about the target's level of military power from the democracy treatment, which is perhaps the most obvious potential confounder." They seem to see a democracy treatment, even one that is randomly assigned, as somehow confounded with an uncontrolled feature. A possible reason is that these authors are not in fact interested in the effect of the treatment itself but of the beliefs that the treatment seeks to change. Dafoe, Zhang, and Caughey (2018) address this issue directly, writing "When IE [information equivalence] is violated, the effect of the manipulation need not correspond to the quantity of interest (*the effect of beliefs about the focal attribute*)" (emphasis added). I return to this below.

The precision argument. The precision argument explains the use of controls at the analysis stage, but the precision argument for adding controls in the intervention stage is less obvious: in the context of a conjoint, adding control during the intervention can *increase* uncertainty, since it can increase variation, which then needs to be removed to get back to baseline in the analysis stage (this idea is developed in section Section D).

The targeting argument. The bigger role that controls play is in *estimand definition*. In a typical experiment, controlling a background condition to some value at the intervention stage suggests an interest in effects *given that (pre-treatment) feature is fixed at that value*. Fisher urges in such cases to vary background conditions using a robustness argument, and such variation is typical in conjoint experiments also. However there are two subtleties that matter a lot for the effects of controls as they are used in survey experiments. First the effect of a given feature depends not just on the *level* of other features, but on whether or not another feature is controlled at all. Second in survey experiments there is sometimes an attempt to control features that are "downstream" with respect to other features, or at least with respect to their real world analogues.

To help fix ideas, Figure 2 shows a simplified model in which I distinguish between attributes (of candidates, say), information provided by researchers about attributes, beliefs subjects have about attributes, and subject decisions, such as their evaluations of a candidate relative to others. Putting concerns from the last section aside we imagine that there is (in the world of forms) a hypothetical candidate with attributes, an experimenter provides information about these attributes, the subject then forms beliefs about these attributes, and then takes an action.

Table 2 then describes a set of estimands researchers might be interested given the set up in Figure 2.

Note the difference between the conditional effects and the controlled effects is that the conditional effects can be thought of as the effect of one treatment given some "background" condition—background here meaning simply that the condition is set at the moment of the application of the experimental treatment as is not the result of treatment. The controlled

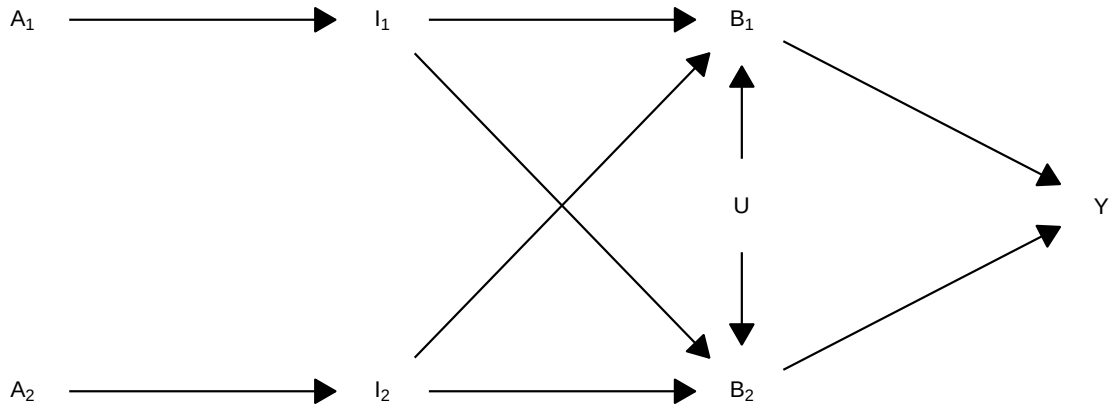

Figure 2: DAG for attributes (A_1, A_2), information (I_1, I_2), beliefs (B_1, B_2), and outcome (Y).

Table 2: Estimands researchers might be interested in given the setup in Figure 2. Last column indicates whether the quantity is identified via randomization of I_1.

	Name	Definition	Short description	Identified?
1	Attribute effect	$Y(A_1 = 1) - Y(A_1 = 0)$	Total effect of attribute A_1 on Y through information and beliefs (including spillovers via B_2).	No
2	Information effect	$Y(I_1 = 1) - Y(I_1 = 0)$	Total effect of information I_1 on Y, via B_1 and induced changes in B_2.	Yes
3	Belief effect	$Y(B_1 = 1) - Y(B_1 = 0)$	Total effect of belief B_1 on Y, allowing correlation with B_2.	No
4	Conditional information effect	$Y(I_1 = 1, I_2) - Y(I_1 = 0, I_2)$	Effect of I_1 on Y given I_2.	Yes
5	Conditional belief effect	$Y(I_1 = 1, I_2) - Y(I_1 = 0, I_2)$	Effect of B_1 on Y given I_2.	No
6	Controlled information effect	$Y(I_1 = 1, B_2) - Y(I_1 = 0, B_2)$	Effect of I_1 on Y, with B_2 held constant.	No
7	Controlled belief effect	$Y(B_1 = 1, B_2) - Y(B_1 = 0, B_2)$	Effect of B_1 on Y, with B_2 held constant.	No

effects however keep a feature constant that is *downstream* to the treatment of interest, here, conditional on a belief that arises as a result of the provision of the signal.

Studies are often described as targeting the first estimand, I discuss versions of this estimand in section Section 4.

The second and fourth estimands are directly targeted by a conjoint experiment. The AMCE is a version of the fourth estimand. For the fourth estimand the role of controlling is not to address confounding but to specify the various conditions that define the conditional effect of interest. For these estimands, the types of confounding that the experiment addresses might include the joint assignment of treatments—for example whenever information is given about democracy, information is also provided about wealth (not simply "*inferences are made about wealth*"), or self-selection into treatment: subjects that are more supportive of a particular type of candidate are more likely to receive one signal than another. Within the context of a survey question it is not obvious why these threats would arise in the first place; nevertheless randomization effectively deals with them. To be clear though the confounding problem addressed here is not the problem that in the world the effect of democracy on alliances is confounded by trading relations. Including controlled attribute signals–for those attribute signals that you want to condition on—is critical to the *definition* of the estimand. The control matters when the effect of one signal likely depends on the presence or the value of other signals. This might be because these add realism because of interaction effects between signals. If one is interested just in the effects of receiving information—perhaps conditional on other signals, given whatever you believe already, then the effect is identified just from random assignment.

Dafoe, Zhang, and Caughey (2018) appear to suggest an interest in estimand 7: "what unifies studies of epistemic effects is their goal of inducing different subjects to consider two alternative versions of a scenario, one in which the factor of interest is present and one in which it is absent, *without affecting subjects' background beliefs*." Background beliefs, in their account, are "[beliefs about] those factors that in the real world are not affected by treatment," by which is meant, presumably, "not affected by real world analogues of the treatment."

The dependence of the interpretation of (implicit) average effects estimands in a factorial experiment on the distribution of other conditions is well appreciated; just as important however is the dependence on which factors are included.[4] Although one might have a notion of fundamental preferences—what would be valued in a full information environment—in settings with imperfect information, preferences over options depend critically on beliefs of unknown features of those options.

3.2 Conditional effects: How estimands depend on attribute sets

It is well appreciated that the effect of one attribute depends on the value of another attribute. This arises whenever the effects of attributes interact. An implication of this fact is that one

[4]By implicitly I mean the estimand that the design targets. Targeting is of course a choice of the experimenter and not something that can be read off from the design.

cannot interpret average effects of one attribute signal without specifying the distribution of other signals, over which you are averaging.

More subtly, perhaps, the effect of one attribute (signal) depends on which attributes *are included in the design* even if attribute signals enter linearly in evaluations.

I illustrate this feature here for a case where the inclusion of an attribute that a subject thinks is causally irrelevant for the quality of an outcome nevertheless dramatically affects the effect of information on the attribute to beliefs about the outcome.

The reason for the interference, despite randomization, is that the effects of beliefs about attributes and beliefs about outcomes does mimic the relationship between beliefs about the effect of attributes on outcomes. Although attributes are randomized in the presentation of profiles, there is no reason to think that *respondents believe they are randomly assigned* or for respondents to believe that candidates are randomly assigned in their mental models of the world.

Consider now a setting where where respondents posit three attributes, ability, privilege, and wealth, that combine to produce "quality" Y. We will take ability to be unmeasured.

Two possible mental models are shown in Figure 3:

- In the first, they think that both privilege and wealth matter, both contribute directly to a quality assessment, alongside ability.
- In the second model they believe that neither privilege or wealth affect quality: ability is the thing that matters. Privilege and wealth each occur independently with probability 0.5. Ability is produced endogenously, where ability = 1 if either privilege = 1 or wealth = 1 (or both).

The first world is a simple world in which each effect can be estimated simply regardless of which other attributes are included: the mental model comports with the inference model. In the second case there is disconnect. In this world we can see citizens do not think that wealth causes quality, but it is absolutely *informative* about quality, since it suggests ability. Privilege however is not (absent information on wealth) informative about quality at all: it neither has an effect nor carries information. Changing the information about privilege changes the beliefs about Y, given the mental model even though *in* the mental model, privilege has no effect on quality.

The intuition is that when they see privilege but no wealth they are confident that the candidate must be low ability. When they see wealth but no privilege they are confident that the candidate must be of high ability. In other cases they are less sure. So, adding a control has generated an "effect" that otherwise would not have been present. If this were observational data we would say that controlling for wealth produces a spurious correlation between privilege and quality. In this context however the inference in not quite the same, rather, conditional on wealth there *is* an effect of information about privilege on beliefs, even though in the voter's, possibly correct, model of the world, privilege has no effect on quality.

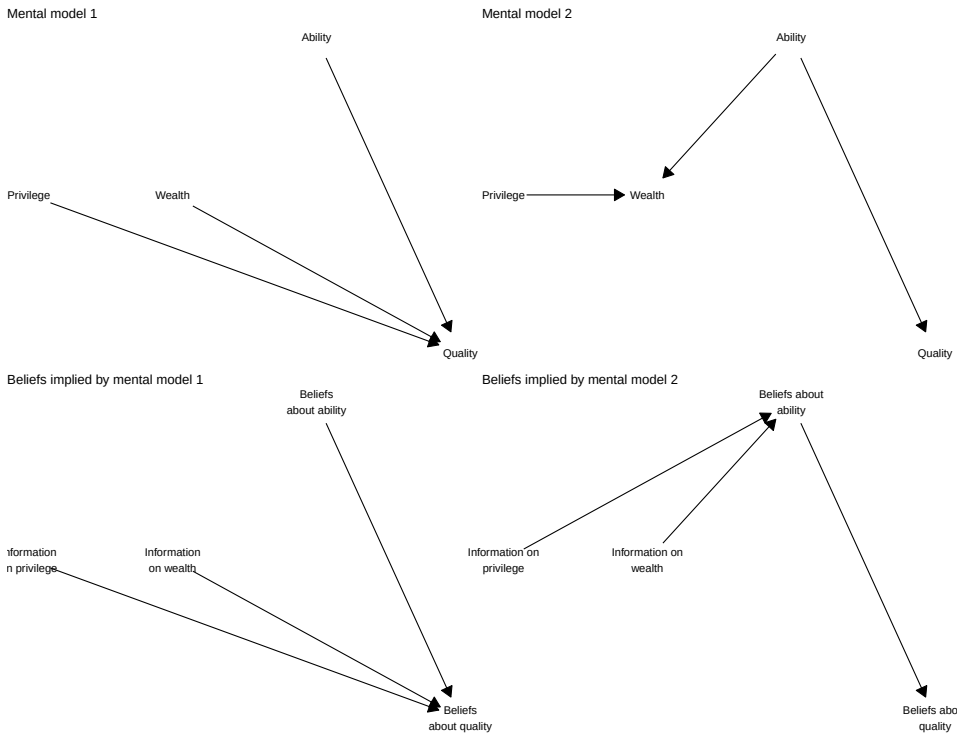

Figure 3: Top row shows three mental models in which three attributes (privilege, wealth, ability) combine in different ways to produce quality. Bottom row shows relations between signals, beliefs, and evaluations, when signals are provided for two of the three attributes only. Signals for an attribute directly affects beliefs about that attribute (not shown), but may also affect beliefs about other attributes for which a signal is not provided. Confounding or collider bias can render a *signal* of a feature relevant to assessments even though—in the individual's mental model of the world—the feature itself does not affect assessments.

From these considerations, under Assumption *, we can extract the implied "preferences" in each world given different control strategies.

Functional equation for Y given:	World 1	World 2
data on A, W, P	$Y = \frac{1}{3} - \frac{1}{3}P + \frac{2}{3}W$	$Y = A$
data on P only	$Y = \frac{2}{3} - \frac{1}{3}P$	$Y = 0.25$
data on W only	$Y = \frac{1}{6} + \frac{2}{3}W$	$Y = \frac{1}{8} + \frac{1}{2}W$
data on P and W	$Y = \frac{1}{3} - \frac{1}{3}P + \frac{2}{3}W$	$Y = \frac{1}{3} - \frac{1}{3}P + \frac{2}{3}W$

We assume here that data on an attribute determines beliefs about the attribute and the equations specify evaluations given beliefs. In the first case with full data, we assume that the equation also captures subject beliefs about the effects of actual attributes on outcomes (this is the realistic Bayesian assumption in Dafoe, Zhang, and Caughey (2018)).

Note that in all cases specifications are linear. What matters is the inclusion or exclusion of the control, not its value.

From the table we see that in the two worlds subjects have starkly different views on what determines a high quality candidate. Indeed under full information they would disagree entirely about what features are relevant. With access to just one feature they would also disagree.

In the first world the marginal effects can be estimated correctly *regardless* of which other variables are included. In the second world the effects of P and W depend entirely on what else is included.

Note that Dafoe, Zhang, and Caughey (2018) engage with a similar logic but come to a different conclusion: whereas I argue here that the estimand changes across settings, they argue that the effect of beliefs are not identified in these models. I discuss the difference in interpretations in Appendix Section E.

Recognizing that conjoints assess the effects of signals on statement, not of attributes themselves, clarifies that the causal map relating treatments to statements does not need to resemble the causal map relating the attributes that are signaled to the properties about which statements are made. Indeed the causal mapping that subjects hold, if any, is not recoverable from the mapping from signals to statements.To see this note that in this example when *both* W and P are included, the induced preference relation is the same in World 1 and World 2. All estimates of marginal effects would also be the same. This example clarifies that one cannot map back from behavior in a conjoint with a fixed set of characteristics to respondent's models of the world with those characteristics. In particular one cannot infer from the fact that information about a feature increases the evaluation of a candidate that the subject believes that the feature itself improves the quality of a candidate. See Section C for further discussion.

A practical implication is that the decision whether to include a control in this context then should plausibly be driven not by confounding concerns, but *on whether the information is*

likely available to the candidate in target applications of interest. The question really: what is the effect of some new information given some prior information set. Kirkland and Coppock (2018)'s study exemplifies the logic nicely: effects of features when partisan information is available are weaker than that when partisan information is not available. Neither set of effects is wrong, necessarily. But one—that with information—may be closer to the target real world estimands.

3.3 Controlled effects and taste-based discrimination

Whereas conjoint experiments are often interested in the effects of some information given background information or background beliefs, in some instances the interest is specifically on controlling for beliefs that are conceptualized as downstream.

An important case is the study of discriminatory behavior or attitudes, in particular to assess how judgments depend on one attribute—race, gender, for instance—while specifically controlling for other features that are plausibly associated with the attribute, at least in the minds of subjects.

Boittin, Fisher, and Mo (2024) suggest that discrimination *after* controlling for criminality suggests taste based discrimination (though they consider other possibilities also). Ono and Burden (2019) argue that controlling for other features lets them assess "taste based" discrimination. Olinger et al. (2024) make a similar argument for their study of doctor selection as a function of race.

This is a productive use of controls where the goal is to block off a set of channels through which a signal operates. It is especially useful when we expect that beliefs about third attributes – like skills – are indeed uncertain and likely to be affected by beliefs about other features of interest, such as identity.

For concreteness imagine that evaluations depend on both gender and skills. In the real world skills might be a function of gender. But gender might also simply be correlated with skills even if it does not cause it. In either case *information* about gender could give rise to *beliefs* about skills. The *controlled* effect of interest—to capture "taste-based" discrimination, is the effect of gender *keeping fixed* the beliefs about skills, that could otherwise change upon learning gender, not simply conditional on background beliefs about skills. The difference is that one might condition on background beliefs about skills, supply information about identity, and find an effect, conditional on background beliefs about skills, but due uniquely to updated beliefs about skills. For taste based discrimination we want to *prevent* beliefs about skills from being updated as identity information is introduced.

Given some form of assumption like Assumption 1 below, exogenous control of downstream beliefs allows for the estimation of the controlled direct effect of the identity cue.

Assumption 1 [perfect belief compliance]: $\sigma_X \neq \emptyset \rightarrow \mu_X = \sigma_X$. Informally, if a signal is provided about an attribute, then downstream beliefs about the attribute correspond to the

signal. The converse is not assumed. The assumption rules out the possibility that beliefs about one attribute are affected by signals on another if the attribute has itself been signaled; but does not rule out cross attribute effects otherwise.

This might be thought of as treating an intervention on information as a "hard" intervention on downstream beliefs. We might have a violation of this assumption if, say, the provision of the interpretation of a *controlled* attribute signal is affected by the presence of another attribute signal. For instance if signal σ_A is "this politician has integrity" and signal σ_B is "this politician supports abortion" then some subjects might reassess their beliefs μ_A that the politician has integrity upon observing the two signals together. Adding signal 2 does not then simply block a channel from μ_A to μ_B, it alters μ_A itself.

In effect this is a mediation problem, and studies of mediation analysis show just how difficult identifying mediation effects are in experimental work (Green, Ha, and Bullock 2010). Yet the problem appears relatively easy, at least for some kinds of mediation estimands, in survey experiments. Acharya, Blackwell, and Sen (2018) give an especially clear treatment.

Two distinct interpretative risks arise in this setting.

First there is risk of interpreting the estimand as "actual" rather than "possible" taste based discrimination. When fixing downstream conditions exogeneously researchers are estimating controlled rather than "natural" estimands. To see the difference, imagine an employer values employees if they are high skilled *or* if they are male. And otherwise do not value them. They also think that everyone is high skilled. For such a person learning that someone is male or female never *in fact* makes a difference (regardless of the actual distribution of skills in the world). They engage in no "taste based" discrimination and so in this simple set up the "natural direct effect" of gender is 0. However the "controlled direct effect" of gender is 1 conditional on low skill. Controlling helps uncover a feature of preferences and of possible discrimination, not actual discrimination. Mapping back to actual discrimination requires information about what the employers would believe under different conditions.

Second there is a risk of inadvertent over control. The fact that downstream controls, in the sense provided in assumption 1 are easy in a survey context raises the risk that controlled effects are returned when conditional effects, or even simple average effect are sought/ Similarly it is possible that estimands are defined with too much control. If in effect we are estimating a set of controlled direct effects then we learn little about average effects. Moreover what we do learn about – direct effects – depends, again, critically on what is controlled for. Controlling for employment or networks when assessing the effects of migration backgrounds, when these things are more normally inferred from information on migration background, then moves the estimand to a controlled effect, remove in effect channels through which information about migration might usually operate. It is not hard to imagine, moreover, that the controlled effect could ultimately be thin soup. In a line of dominos, each having a unit effect on the fall of the final domino, only the second last domino has a non zero controlled direct effect. For a type of reductio, if we imagine that the effect of any attribute works via beliefs about downstream features, then by controlling these features we can send the effects of the attribute to zero.

If we think "taste" effects work through particular channels: disgust, affection, beliefs about integrity or piousness—then the ability to control for these (normally downstream) mediators would remove direct effects entirely.

4 Mapping from estimates from survey experiments to real world estimands

Sometimes survey experiments are implemented not because researchers are primarily interested in public opinion, but rather are interested in using public opinion to understand causal processes in the world. For instance a prime reminding people of social identities, relative income, or histories of injustice are used not because of an interest in the effect of primes, but because of an interest in the effects of social identities, relative income, or histories of injustice. A vignette experiment varies a description of a state as democratic not simply to understand how statements of public attitudes depend on democratic descriptors but to understand whether being a democracy matters for escalation risks. Outside of the survey setting, a researcher might use an audit experiment, not to understand how employers would respond to different types of CVs, but whether people fare better or worse because of their social identities.

These all require making inferences from estimands identified within the context of a survey to real-world estimands.

In the case of conjoints in particular researchers often appear at least to draw quite strong conclusions about real world processes on the basis of findings from survey experiments. Kertzer, Renshon, and Yarhi-Milo (2021) describe, for instance, how "allies who stood firm in the past indeed gain a reputation for resolve and are seen as more likely to stand firm in the current crisis." Mares and Visconti (2020) report how the "promise to renovate schools increases the probability of support by only four percentage points (over candidates who do not make these promises)." Tomz and Weeks (2013) report how "shared democracy pacifies the public primarily by changing perceptions of threat and morality." Amsalem and Zoizner (2024) describe their study as capturing the causal effect of candidate extremity on citizens' preferences.

Such claims might be read as if they are statements about the effects of features of the world on real outcomes, though they are statements about the effects of *controlled* changes in *descriptions* of hypothetical packages on imagined responses. They can sometimes sound as if candidates and not voters are being treated.

In the same way, there is a risk of misreading Bansak et al. (2023)'s claim that the ACME is a "tool for analyzing elections" that can be used to estimate "the effect of a change in an attribute on a candidate's or party's expected vote share," to think that they are interested in the effect of a candidate's attribute on the candidate's vote share.[5]

[5]Interestingly the statement of Proposition 2 in Bansak et al. (2023) is not described in terms of 'effects' of an attribute but simply differences between vote shares of candidates with different attribute values. The

Can one move between one type of statement and the other?

These statements are interesting for two reasons. First they seek to move from inferences about estimands generated under controlled conditions to inferences about estimands generated "observationally." Second these statements are interesting cases of a "micro-macro linkage" problem (Humphreys and Scacco 2020). The returned vote share for a candidate is a macro quantity, possibly affected by the attributes of a candidate. But it is generated from the individual preferences of a multitude of voters, each based (at least in part) on the information they have, which is likely itself a function of candidate attributes. Indeed this seems like it should be a particularly simple micro-macro linkage problem since there is a well defined aggregation rule that one should be able to plug in to make the link.

In this section I make two arguments. First that the "observational estimands" to which experimenters seek to make inferences may themselves not be well defined, making the attempt to map foolhardy. Second even if there are well defined mapped estimands, the conditions required to make inferences from one estimand to the other may be extremely difficult to satisfy.

4.1 Maps to nowhere

Conjoint experiments have little problem estimating the effects of *signal* about gender, about race, about migration status, about democracy, about all kinds of features. The effect of a issuing a *signal* on a response is well defined. But the corresponding estimands – the effects of a candidate's gender, of a country's regime type, and so on, may not be.

The difficulty relates to a curious point of slippage in the extension of methods of causal inference to observational studies. As described by Rubin (2005), much of the early work on causal inference took place within the context of treatments of experiments in which research controlled the assignment of treatments. Contributions by Rubin, Pearl (2009), and others extended the scope to the study of non-experimental processes. These processes differ in two ways, one well appreciated, the other less well appreciated. The first is that thanks to control of assignment to treatment, researchers can avoid confounding and ensure that effects are identified. The second is more fundamental: *since* they control treatments, experimentalist usually have a well-defined treatment, with —typically— well-defined estimands.

In contrast, observationalists face twin problems: the well-appreciated problem of ensuring they can estimate estimands and the prior, less often discussed problem, of ensuring that their estimands are well defined. This, I think, is particularly true for the effects of features, or states, rather than the effects of interventions. Conjoints are very often concerned with features rather than interventions and the causal interpretation requires a notion of interventions that alter these features. "Surgery on equations" in Pearl's formulation. Manipulation in Holland's.

text before and after the proposition interprets the results as licence for causal inferences about the effects of attributes.

There are three threats to the meaningfulness of observational causal estimands that are defined over features rather than with respect to particular specified interventions.

Attributes as causes. Holland argues most strongly that "attributes of units are never causes." His thinking is that if you change the attribute you are no longer talking about the same unit. Though he describes attributes he seems to have in mind features that are constitutive of the unit. Would circles have such a low perimeter to area ratio if they had corners? Or to modify an example in Rubin (2005), would Trump have won the election if he were born yesterday in the Arctic). More conservatively, Holland interprets scores on tests as an attribute and challenges causal interpretations of statements of the form "She did very well on the exam because she studied for it." In the same way one might argue that refusing to contemplate alternative genders or racial identities for the same person is overly restrictive. Features should not be constitutive of the unit.

SUTVA violations. The second threat is, I think, more common. The second component of the "stable unit treatment value assumption" (SUTVA) is that there are no hidden versions of treatments. What's the effect of having an even number of parties taking part in government negotiations? The problem here is that there are many different ways that the treatment condition could be met, with different implications. We might specify multiple versions and put a distribution over them (VanderWeele (2018); VanderWeele and Hernan (2013)), but absent such measures we are in inferential trouble. The problem of underspecification is especially acute when we think about states rather than interventions. States certainly have effects. The fact that my bicycle *has* a puncture matters. But implicitly when we talk about states we are talking about states with time stamps: my bicycle was punctured when I went to ride it; whether it was punctured the day before is irrelevant. What is the effect of "being a democracy" requires not just a notion of what not being a democracy is but also a specification of *when* the unit *was* a democracy compared to the situation had it not been a democracy at that time. Yet this temporal feature is not specified and we can imagine many versions of the treatment meeting the definition: a state that has been a democracy for 200 years, or one that became a democracy yesterday. Similarly we can imagine many versions of what it means to be a migrant, being employed, or being wealthy, all with distinct potential outcomes.

Exclusion restriction. The third challenge is if the levers we can imagine to modify a feature inevitably induce effects of their own on outcomes. The difficulty contemplating the effect of a change in one feature without necessarily inducing a change in another. In the case of my punctured bicycle I can imagine outcomes with the puncture fixed, whether I fixed it or it mended itself somehow. But let's imagine that we are interested in a candidate's gender or the democratic state of a country. Can we imagine a (timestamped) change in democracy without imagining that there was a revolution of some form? Can we imagine a change to Trump's gender without imagining that he transitioned somehow. If not, and it seems definitionally we cannot, we have to accept that the gender effect includes the effect of a gender transition, though this is surely not what we have in mind.

Although these ideas are abstract, the key implication is simple. Not everything is a cause, but statements about anything could be a cause. We err though if we think that the measurement

of causal effects of statements about a feature imply that the features themselves have causal effects (much less what those effects are).

4.2 Licence to export

Let's turn now to the problem of making inferences about real world estimands, assuming now that these are indeed well defined. Imagine a "target" setting in which we have an actual election with a large set of potential candidates in two candidate elections who are defined by a finite set of attributes. Imagine we also have a conjoint experiment that closely parallels the setting.

In particular, survey participants are representative of voters, the set of attributes of candidates that matters to voters is known and signals about this set are presented to survey participants, who react to the signals provided in the survey in the same way as they would do as voters to the information on attributes of candidates in the election.

Can we learn about the effect of the attribute on vote choice from the conjoint? More specifically we will ask whether the AMCE of the signal of attribute (for example, a gender signal) can correspond to the ATE of that attribute on votes in an election, where both the AMCE and the ATE are defined with respect to the same population of voters and the population of potential candidates (so to be clear we are not focused on the effect of gender on vote shares given the actual candidates running in a race, but the expected effect across races that might have been run).[6]

We make the connection by assuming that signals about attributes (s) in the world are possible functions of attributes (a); that voters preferences over candidates, (Y) depend on the signals they receive about candidates; and that preferences in turn translate into votes (V).

Specifically, consider an election between candidates j and k in context u. Context u induces a set of attributes for candidates—as measured at election time, with a denoting a vector of all attributes for both candidates.

Let us say that the vote share received by a candidate depends only on the attributes of the two candidates: $v_j(a)$.

We are interested in the effect of a particular attribute of one candidate, which we take to be the first element of a, a_1, on the vote share of candidate 1 in context u. Let v_0^u and v_1^u denote the potential outcomes (vote shares) for candidate 1 when attribute a_1 is set to 0 or 1, respectively. Let $y_i^u(s^i) \in \{0, 1\}$ denote whether voter i prefers candidate 1 to candidate 2 given a vector of signals s^i in context u. Let $a_{-1}(a_1, u)$ denote the potential values of other attributes following a notional intervention on attribute a_1 in context u.

[6]The exercise is related to Bansak et al. (2023)'s interest in settings with the "election matching the specifications of the conjoint," though they do not have an external estimand in mind. They are not attempting to map to the effect of a real candidate's actual attributes on actual vote shares and their Proposition 2 does not seek to show an equivalence between two different estimands.

4.2.1 The ATE

Let β^u denote the candidate-level effect of the attribute on the vote share in context u:

$$\beta^u \equiv v_1^u - v_0^u.$$

Allowing for the possibility that a controlled change in one attribute induces endogenous changes in other attributes, we may write

$$\beta^u \equiv v(1, a_{-1}(1,u), u) - v(0, a_{-1}(0,u), u),$$

where $a_{-1}^u(z)$ denotes the potential outcomes of all other attributes when attribute a_1 is set to $z \in \{0,1\}$ in context u.

It is worth highlighting that we assume these potential outcomes are well defined. This presupposes that a change in attributes does not alter the fact that the election goes ahead with two candidates and produces votes for the two candidates (for example, one might imagine a counterfactual attribute under which one of the candidates has an authoritarian agenda).

We now consider the average effect of the attribute across a range of possible contexts, where u is distributed according to h:

$$\beta \equiv \mathbb{E}_h[v^u(1, a_{-1}^u(1)) - v^u(0, a_{-1}^u(0))]\,.$$

In practice the set of contexts one might be interested might be very narrow, for example conditioning on all features of a given election and so imagining only counterfactual values on a_1; in which case for some u, $\beta = \beta^u$. Thus the meaning of β, even whether it is an average effect, depends on h.

4.2.2 The AMCE

Let $i \in N$ have an information vector s^i on candidates j and k and state a preference y_i for j over k. We think of s^i as the signals that i receives about a. However, the elements of s need not stand in one-to-one correspondence with the attribute vectors a—for instance we might imagine a has much larger dimensionality than s.

We do assume however that s_1 is associated with a_1 in the minimal sense that we will want to compare the effect of a change in s_1 to a change in a_1—we do not (yet) assume that s_1 is informative about a_1.

The effect of one element of the information set s^i—say the gender of candidate j—on stated preferences, holding all other information fixed, can be written as

$$\tau_i^u \equiv y_i(1, s_{-1}^i, u) - y_i(0, s_{-1}^i, u),$$

where s_{-1}^i denotes all other information on both candidates.

Letting μ denote the population distribution over individuals $i \in N$, the average treatment effect for s_1, holding other information fixed, is

$$\tau^u \equiv \mathbb{E}_\mu\left[y_i(1, s_{-1}^i(u), u) - y_i(0, s_{-1}^i(u), u)\right].$$

If we are interested in the expectation of this treatment effect under a distribution of signals, f, specified by the researcher, in a particular setting u, we have:

$$\text{AMCE}^{f,u} \equiv \mathbb{E}_f\left[\mathbb{E}_\mu\left[y_i(1, s_{-1}^i, u) - y_i(0, s_{-1}^i, u)\right]\right].$$

A focus on AMCE rather than τ^s is appropriate when interest lies not only in the effect of an attribute under a given information environment (such as that induced by a particular experiment), but in the average effect across a range of possible informational settings. The distribution of signals here is as determined by h, which is what a researcher might attempt to replicate in an experiment.

Note that τ^u and the AMCE here are written as controlled effects. Two other more "natural" versions might be of the form:

$$\text{AMCE} \equiv \mathbb{E}_h\left[\mathbb{E}_\mu\left[y_i(1, s_{-1}^i(u), u) - y_i(0, s_{-1}^i(u), u)\right]\right].$$

here the signals are again controlled, but at the levels they occur naturally at in the given context.

However, if we think of s^i as denoting *background* features for an individual i and with the usual exclusion restriction under which an intervention on s_1 does not affect other background features, these are equivalent to the familiar unit level treatment effect $y_i(1) - y_i(0)$ (suppressing the background information) and similarly the ACME can be described with the familiar expression $\tau = \mathbb{E}[Y(1) - Y(0)]$, where we use $Y(.) \equiv y_i(.)$ with $i \sim \mu$ to denote the random potential outcome induced from sampling i from N.

4.2.3 Equivalence

The question is whether, there are conditions under which we have:

$$\beta = \text{AMCE}.$$

Note the question is about the equivalence of estimands, not about whether either estimand can be effectively estimated. Proposition 1 below provides a positive answer under the following conditions.

A1. Causal autonomy of attributes.
For all contexts u and all g,

$$a^u_{-1}(a_1 = 1) = a^u_{-1}(a_1 = 0).$$

Note: Intuitively, an intervention on an attribute does not causally affect other attributes. Attributes may of course be arbitrarily correlated with each other.

A2. Sovereignty.
Vote shares are determined by votes only. Let $v_j(a)$ denote the vote share received by candidate j. Then

$$v_j(a) = \mathbb{E}_\mu[v^j_i(a)].$$

Note: Implicitly, votes translate directly into vote shares. This is trivial if understood definitionally. However if v is understood as the *reported* vote share then this may be considered a "no misreporting" condition.

A3. Sincere (rational, but nonstrategic) voting.
Let $y_i(a)$ indicate whether candidate 1 is i's preferred candidate given attribute vector a, and let $v_i(a)$ indicate whether i votes for j.

$$v_i(a) = y_i(a).$$

Note: Implicitly: voters do not abstain—regardless of a, they vote their top preferences. Rational but nonstrategic voting may sound a little oxymoronic, it captures however the idea that they vote their top preferences but do not take into account any other considerations, such as the likely voting behavior of others.

A4. Context irrelevance given attributes.
For all contexts u and attribute vectors a, for all u, u':

$$y_i(a, u) = y_i(a, u').$$

Note: In particular features that give rise to different attributes—one might imagine, cultural features for instance—do not themselves determine preferences.

A5. Signals are complete mediators.
For all individuals i,

$$y_i(1, a_{-1}) - y_i(0, a_{-1}) = y_i(1, s^i_{-1}) - y_i(0, s^i_{-1}).$$

We typically might expect voters have fundamental preferences are over attributes and they use signals to infer attributes. As a simple example a utility maximizing voter would have:

$$y_i = \begin{cases} 1 & \text{if } \sum_a q^{ij}(a \mid s^i) u_i(a) \geq \sum_a q^{ik}(a \mid s^i) u_i(a), \\ 0 & \text{otherwise.} \end{cases}$$

where $q^i(a|s^i)$ denotes the *probability distribution* that i places over possible attribute vectors $a \in A$, given signal s_i.

Under this model this axiom might be satisfied if s and a have the same domain, and for all attributes $s(a) = a$. In this case Bayesian voters infer a from $s(a)$ without error and

$$y_i = \begin{cases} 1 & \text{if } u^j_i(a) \geq u^k_i(a), \\ 0 & \text{otherwise.} \end{cases}$$

A6. Aligned distributions.
The distribution of signals induced by contexts corresponds to f:

$$\mathbb{E}_h\left[\mathbb{E}_\mu[y_i(s^{u,i})]\right] = \mathbb{E}_f\left[\mathbb{E}_\mu[y_i(s^i)]\right].$$

Claim. *Given a population of voters $N = [0, 1]$, if conditions A1–A6 hold, then the expected effect of a change in an attribute over contexts on vote shares of a candidate (β) equals the expected effect of a change in signals of that attribute on expressed preferences, conditional on other information (AMCE).*

Proof. We have a direct proof in six steps:

$$\begin{aligned}
\beta &= \mathbb{E}_h[v^u(1, a^u_{-1}(1)) - v^u(0, a^u_{-1}(0))] \\
&= \mathbb{E}_h[v^u(1, a^u_{-1}) - v^u(0, a^u_{-1})] && \text{(A1: Causal autonomy of attributes)} \\
&= \mathbb{E}_h\left[\mathbb{E}_\mu[v^u_i(1, a^u_{-1}) - v^u_i(0, a^u_{-1})]\right] && \text{(A2: Sovereignty)} \\
&= \mathbb{E}_h\left[\mathbb{E}_\mu[y^u_i(1, a^u_{-1}) - y^u_i(0, a^u_{-1})]\right] && \text{(A3: Rational non-strategic voting)} \\
&= \mathbb{E}_h\left[\mathbb{E}_\mu[y_i(1, a^u_{-1}) - y_i(0, a^u_{-1})]\right] && \text{(A4: Context irrelevance given attributes)} \\
&= \mathbb{E}_h\left[\mathbb{E}_\mu\left[y_i(1, s^{u,i}_{-1}) - y_i(0, s^{u,i}_{-1})\right]\right] && \text{(A5: Signals are complete mediators)} \\
&= \mathbb{E}_f\left[\mathbb{E}_\mu[y_i(1, s^i_{-1}) - y_i(0, s^i_{-1})]\right] && \text{(A6: Aligned distributions)} \\
&= \text{AMCE.}
\end{aligned}$$

□

4.3 Three challenges

For an example of a structure under which equivalence might hold, consider a causal model like that shown in Figure 4.

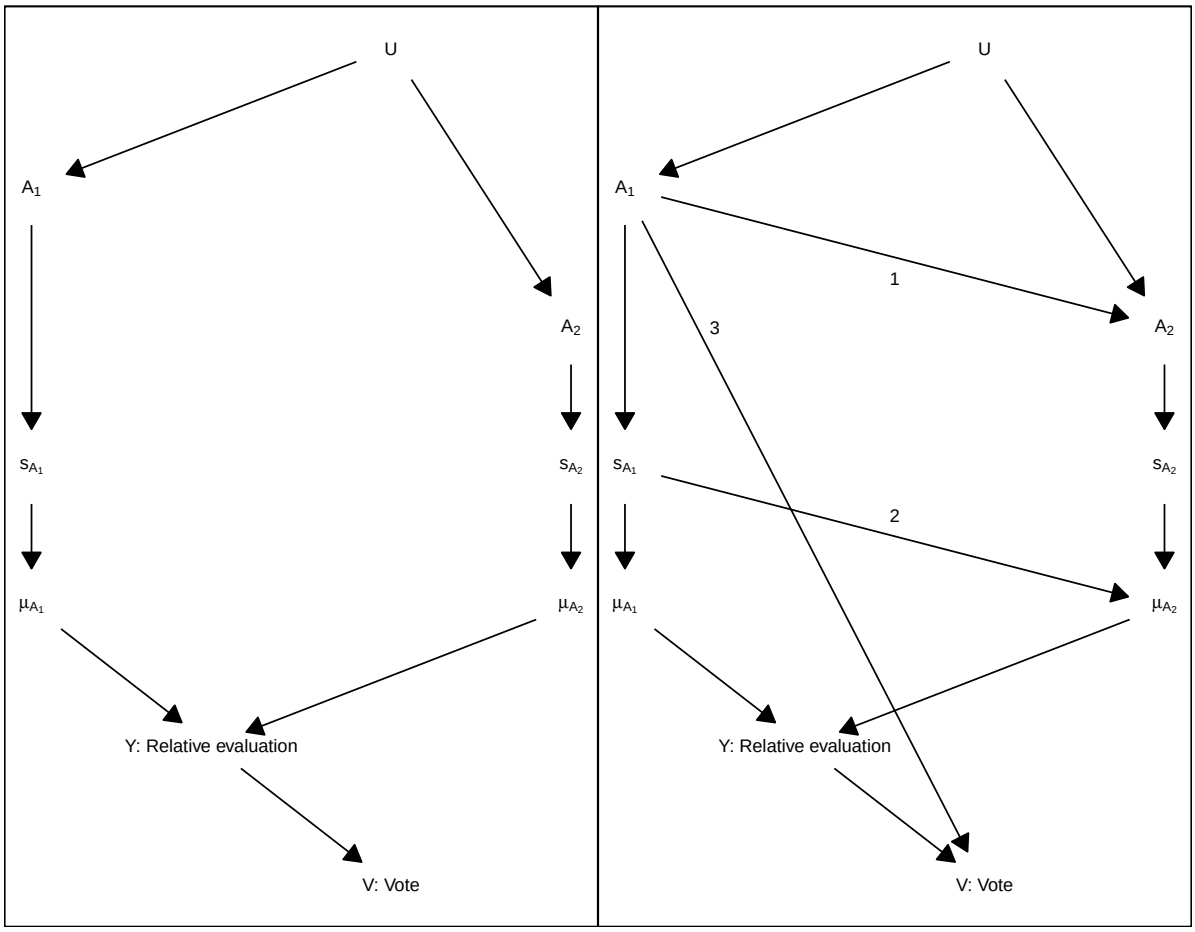

Figure 4: DAG showing relation between two features, A_1 and A_2, an unobservable feature U, two *signals* about these observable features s_1 and s_2, respondent *beliefs* about these features μ_1 and μ_2, and a respondent level (relative) evaluation, Y, and a vote choice. A survey experimentalist intervenes on s_1 and s_2. Can they make inferences about the effect of A_1 and A_2 on V? Left model shows a promising case. Right model highlights a set of threats.

The assumed structure on left is a very simple one. The candidate attributes are confounded:

that is that there is an unobserved feature U that simultaneously affects the two attributes A_1 and A_2. However, we imagine that signals are produced only by the features they signal. Beliefs are a function only of signals. Beliefs about attributes—and only these—are what determines evaluations (say relative to some comparison candidate), and evaluations—and only evaluations—give rise to actual vote choices.

From this graph one can see that the confounding means that one cannot assess the effect of A_1 or A_2 on V by looking at simple differences in means, since in each case there is a backdoor path to V via U. Randomizing A_1 and V is not possible. Conditioning on attributes would work however but only if we had information on *all* attributes that open backdoor pathways.

Imagine now that we can control attribute signals; meaning that we can produce *the same signals about candidate features that voters would obtain in the wild* (in the terminology used in Barabas and Jerit (2010) we can deliver the "natural treatment"—the signal as it would be provided in the world).

Consider now three threats to these inferences.

1. Causal relations between attributes
2. Cross attribute updating
3. Direct behavioral effects

Each of these threats is captured by a marked arrow in the modified DAG below.

4.3.1 Mutually endogenous attributes

Although in a conjoint we can ensure that there are no causal relations between *signals* we provide, we do not have the ability to ensure that there are no causal relations between actual attributes in our target application.

From the graph one can see that this matters because the effect of a change in A_1 is not captured anymore by the effect of a change in s_1, even if we allow for A_1 and A_2 to be arbitrarily correlated.

This feature of endogenous relations between attributes seems to raise a tremendous obstacle to making claims from conjoints about the effects of attributes in the world.

To illustrate the effects of a violation of A1, let's assume that A2 - A6 hold. In addition lets optimistically assume both of two attributes for two candidates are observed, all voters have identical preferences (this simplifies the estimand and makes it clearer what variation is doing work), and we know not only the joint distribution of attributes and the average effects of all attributes on each other, but also the underlying response functions giving rise to these average effects (this simplifies calculations of the ATE)

Now imagine two worlds, similar to the simple model in Figure 4. They differ in that in one world A_1 causes A_2 and in the other A_2 causes A_1. All nodes binary.

Let's assume that these two worlds can have *exactly the same observed joint distribution* of candidate attributes. For example, say in both worlds the correlation structure is:

$$\Pr(A_1 = 1) = \Pr(A_2 = 1) = 0.5$$

$$\Pr(A_1 = 1 \mid A_2 = 1) = \Pr(A_1 = 0 \mid A_2 = 0) = 0.8$$

However, in **World 1**, intervening on A_1 induces changes in A_2; in **World 2**, it does not (whereas in a conjoint experiment, when we alter A_1 in a profile description, we do *not* induce changes in A_2, even if A_1 and A_2 are correlated in the population).

Specifically we will assume that in world 1 $A_2(A_1) = A_1$ for 60% of units; for 20%, $A_2 = 0$ regardless of A_1 and for the remaining 20%, $A_2 = 1$ regardless of A_1. Similarly for World 2, $A_1(A_2) = A_2$ for 60% of units; for 20%, $A_1 = 0$ regardless of A_2 and for the remaining 20%, $A_1 = 1$ regardless of A_2. In World 1 it is easy to see that the effect of A_1 on Y arises not just because of the effect of A_1 on beliefs about A_1 but because of the effects of A_1 on A_2 and beliefs about A_2.

Imagine that everyone has very simple preferences across candidates:

$$U(A_1, A_2) = A_2.$$

So voters ignore A_1 entirely. Given two sets of attributes (A_1, A_2) and (A_1', A_2'), the choice

$$Y(A_1, A_2, A_1', A_2')$$

is determined solely by which profile has the larger value of A_2. Ties are broken at random.

We will want to compare the AMCE and the ATE. For the AMCE we have:

$$\text{AMCE}_1 = \sum_{A_2, A_1', A_2'} \Big[Y(1, A_2, A_1', A_2') - Y(0, A_2, A_1', A_2') \Big] \, p(A_2, A_1', A_2').$$

where $p(A_2, A_1', A_2')$ is the joint distribution of attributes for candidates 1 and 2 (other than A_1).

The ATE is

$$\text{ATE}_1 = \mathbb{E} \left[\sum_{A_1', A_2'} \Big[Y(1, A_2(1), A_1', A_2') - Y(0, A_2(0), A_1', A_2') \Big] \, p(A_1', A_2') \right].$$

where the expectation is taken over responses of A_2 to A_1—that is, across candidates, not across voters. In the case of a single candidate of interest there is no need for this expectation.

There is no expectation across voters in either expression because I have simplified things by assuming all voters are identical.

The AMCE is computed by comparing changes in A_1 for candidate 1, holding A_2 fixed, across all possible profiles of candidate 2. In this case, regardless of the values of A_2 or of the rival's attributes, the effect of A_1 is 0 and so the AMCE is 0 (regardless of the correlation structure p).

$$\text{AMCE}_1 = 0.$$

A similar calculation (not shown) yields

$$\text{AMCE}_2 = 0.5.$$

This arises because, depending on the other candidate's attribute, a change in A_2 either takes you from a loss to a toss up or from a toss up to a win.

How about the ATE?

In World 1, changing A_1 may induce changes in A_2. Taking the distribution of profiles for candidate 2 as given, a change in A_1 produces a change in the probability of support of 0.3 *regardless* of the other candidates' attributes. This is the effect of A_1 on A_2 times the effect of A_2 on Y. Table 14 in the appendix shows these calculations explicitly.

In World 2, changing A_1 does not affect A_2. The resulting comparisons show that $\text{ATE}_1^{\text{World 2}} = 0$.

Hence the AMCE coincides with the ATE in World 2 but not in World 1. The difference arises because, in World 1, changing one attribute causally affects another attribute that voters care about.

4.3.2 Cross signal updating

We now do not impose that beliefs for a given feature are a function of information on that feature only, reflecting the possibility that individuals can update using all information. Particularly salient in this context is the possibility that when no signal is provided about some feature, individuals might nevertheless update on that feature based on information on other features. This is an old idea, found in Anderson (1971) and Von Winterfeldt and Fischer (1975) for instance and follows from standard Bayesian logic: if I learn about your profession but not your gender I may nevertheless update on both.

Acharya, Blackwell, and Sen (2018) discuss this possibility in the context of conjoint experiments, highlighting how updating on other features can be part of the mechanism through which a signal of a given feature operates (see also Dafoe, Zhang, and Caughey (2018) on "information equivalence"; as well as discussion of this issue in the context of the use of names in audit type experiments Landgrave and Weller (2022) and recent work by Abramson and Gillespie (nd)).

4.3.3 Direct behavioral effects

Finally it is easy to see that attributes can have effects on behavior that arises not just through preferences but through other channels. For instance a powerful candidate might be more likely to get votes because they can intimidate voters or because they can deliver material benefits. Of course a conjoint might measure intended behavior and not just preferences; this is not resolve the problem since attributes might drive a wedge between anticipated behavior and actual behavior, just as it could do between preferences and behavior.

5 Conclusion

The credibility revolution drew attention to just how hard it is to draw causal inferences about social processes—for example, the effects of democracy on conflict or the effects of institutions on growth. Greater clarity about the challenges of causal inference also shed light on what can or cannot reasonably be considered a cause. As Holland (1986) argued, the "experimental model eliminates many things from being causes, and this is probably very good." It left us conscious of the difficulty of defining some causal effects and the difficulty estimating effects if indeed they are well designed.

The rise of survey experiments seems, at least at first glance, to promise solutions to both difficulties raised by the credibility revolution. Democracy cannot be randomized in the real world, but it can be randomized in a survey experiment. The effects of gender may not be well defined, but the effects of providing information about gender in a survey experiment certainly are.

But, of course, survey experiments do not "solve" these problems. Rather they point our attention to different problems that can be more readily solved. A survey cannot randomize democracy, only hypothetical questions about democracy—or, at best, beliefs about democracy. A survey does not record effects on vote shares, or even on votes themselves. At best, it gathers measures of hypothetical votes or guesses about votes.

These inferential traps are not unique to survey experiments at all, or to conjoints in particular. Some of these concerns arise naturally in audit experiments—such as those studied in Fang, Guess, and Humphreys (2018)—in which myriad features of fictional individuals are assigned

independently. Others arise in information experiments in which information is provided about real individuals but in highly constrained informational settings.

From this discussion, we can gather a few prescriptions.

First, be clear about the nature of your estimand. If your estimand is fundamentally descriptive, assess whether an experimental procedure is the most efficient way to make inferences.

Second, be clear about what the controls are doing. The choice of which controls to include—not simply the values of controls—can directly affect your estimand, whether you are interested in the effects of information, the effects of belief, or simply the preferences induced by different information environments. This has implications not only for which attributes should be included, but also for the interpretation of estimands. Statements of effects of direct effects, such as claims about taste based discrimination should not be described as average effects (as is the case for instance when Olinger et al. (2024) claim "Americans do not select their doctors based on race") but specifically with regard to the set of downstream outcomes that are controlled for. In a similar way we should avoid describing effects without reference to the controlled conditions under which they are defined. And we should pool at our peril.

Finally, keep claims in line with design. In particular, avoid language that confuses attributes of candidates with information provided to subjects, and avoid suggesting that hypothetical candidates are treated when, at best, subjects are. Estimates may shed light on different types of discrimination without implying evidence on the effects of "being" a member of one group rather than another. Avoid claiming effects on choices when, at best, we observe effects on how subjects answer surveys.

A Descriptive and causal estimands for other types of survey experiment

A.1 Priming experiments

Priming experiments can be used for making inferences about both descriptive and causal estimands.

A.1.1 A priming experiment conducted for descriptive inference.

Say I am interested in whether you know (K) that a weapon was used in a crime. Your knowledge is something I think you have or do not have and I want to know about it. I would love to just measure that evidence, but it is hard.

So I show you a picture (X) of the weapon and I measure your reaction (Y). I make inferences about the effect of the prime on your reaction (X on Y) in order to make inferences about your knowledge (K). The effect estimate is a diagnostic tool. I make a causal inference in order to do descriptive inference. But I am clear: my interest is in your knowledge, it is not on the effect of seeing a weapon on your stress levels.

One implication of this is that I would be unhappy with this study if I found no evidence for a causal effect but in fact $K = 1$, or if I did find evidence but $K = 0$; for the simple reason that my interest in the causal effect is just instrumental here.

A.1.2 A priming experiment conducted for causal inference.

But I might well be interested in a priming experiment specifically to make causal inferences. I am interested in whether being reminded of corruption by a politician makes you more likely to support the opposing party. I am interested in this because I think politicians or the media do this before elections and I am interested in understanding these effects. If the focus is on the effect of the prime itself this is a standard causal estimand inferred using an experiment, that may or may not just happen to be delivered using a survey.

That makes lots of sense in principle. In practice I think sometimes we see people can trip up and mix up the effect of the prime (e.g. from being reminded that there is corruption) with the effect of the thing being primed (e.g. the effect of corruption itself), or not be clear on whether in fact new information is being provided or not.

A.2 List experiments

List experiments might also be done for either reason, but the typical use is for descriptive inference.

A.2.1 A list experiment conducted for descriptive inference.

You are interested in whether people think there is corruption (K) or not. In principle this is measurable, but it is hard to measure. You vary whether there is a long list or short list (X) and infer from the effects on the count answers (Y) whether people think there is corruption or not. You are primarily interested in K; there is no independent interest here in how list length affects answer except for its role for descriptive inference.

A.2.2 A list experiment conducted for causal inference.

I think this is not so common but you could imagine being interested in the effect of a long versus short list on whether people exhibit social desirability bias. Here you are interested in the effect of the length itself, or of the mention of the word itself. Blair and Imai (2012), when describing conditions for valid inference of the descriptive estimand, describe a "no design effects" condition that rules out various causal effects. One could in principle be interested in just these (and estimate well if you independently have knowledge of the descriptive estimand)!

There is a good literature comparing experimental and direct approaches for asking sensitive questions. The fact that the estimand is the same in both cases highlights that the focus is typically descriptive. The gain from using an experiment is (hopefully) unbiasedness that comes from providing protection to subjects that require plausible deniability. But the fact that it is an experiment itself implies a cost: you get error from the need to do inference (as well as from complexity; Kuhn and Vivyan (2022)) and so need to determine whether the added error is worth it.

A.3 Framing experiments

Framing experiments typically involve changing of question wording to assess how the way a question is worded affects how subjects respond. This sounds similar to assessing the effect of the question on the answer, but actually something a little more subtle is going on in these experiments. Implicitly there is a distinction between the substance of the question — what is being asked — and the form of the question—how it is asked; and the experiment lets one assess whether the form affects the answer, whether there is a violation of "description invariance" (McKenzie et al. 2025). Indeed the idea of an equivalence frame (Tversky and Kahneman 1987; Druckman 2001) is that the *same* question is asked in different ways. Contrast this to a conjoint where under different treatments different questions are asked (in the same way).

A.3.1 For descriptive inference

Goldin and Reck (2015) (see also Goldin and Reck (2020)) treat framing effects as more of a nuisance than a quantity of interest and seek methods to purge estimates of presumed stable

quantities of interest from framing effects, identifying how using a framing experiment lets one point identify quantities of interest specifically for subgroups who are unaffected by frames.

A.3.2 For causal inference

Many studies examining framing effects are specifically interested in causal quantities, specifically the effects of various types of triggers on the ways people think about issues. The basic example of a loss versus gains framing described in Druckman (2001) is naturally thought of as a causal effect on preferences over a given set of alternatives: the frame affects how one thinks. Thus Sniderman and Piazza (1993), for instance, assess how presenting a redistributive frame (affirmative action) affects the expression of racial animosity.[7]

B Calculations

B.1 World I: Simple calculations

The distributions of nodes they expect to see are as in the table below.

Privilege	Ability	Wealth	Quality
0	0	0	$\frac{1}{3}$
0	0	1	1
0	1	0	$\frac{1}{3}$
0	1	1	1
1	0	0	0
1	0	1	$\frac{2}{3}$
1	1	0	0
1	1	1	$\frac{2}{3}$

Their inference on observing privilege and wealth are then:

Privilege	Wealth	Quality	p
0	0	1/3	1/4
0	1	1	1/4
1	0	0	1/4
1	1	2/3	1/4

[7] Sniderman and Piazza (1993) in fact write that the mention of affirmative action would encourage a dislike of blacks, which presupposes that racial preferences themselves, rather than the expression of preferences, are affected.

Their inference on observing privilege are then:

Privilege	Wealth	Quality	p
0	.	2/3	1/2
1	.	1/3	1/2

Their inference on observing wealth are:

Privilege	Wealth	Quality	p
.	0	1/6	1/2
.	1	5/6	1/2

B.2 World II: Collider bias calculations

The distributions of nodes they expect to see are as in the table below.

Privilege	Ability	Wealth	Quality
0	0	0	0
0	0	0	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	0
1	1	1	1
1	1	1	1

Their inference on observing privilege and wealth are then:

Privilege	Wealth	Quality	p
0	0	1/3	3/8
0	1	1	1/8
1	0	0	1/8
1	1	2/3	3/8

Their inference given P only:

Privilege	Wealth	Quality	p
0	.	1/4	1/2
1	.	1/4	1/2

Their inference given W only:

Privilege	Wealth	Quality	p
.	0	1/8	1/2
.	1	3/8	1/2

Privilege	Wealth	Quality	p
0	0	1/3	3/8
0	1	1	1/8
1	0	0	1/8
1	1	2/3	3/8

The equations in tables X simply summarizes these conditional distributions.

C Control dependent beliefs

Here I give a simple example where the value effect of an attribute A_1 on evaluation Y is either 0, .25 or -.25 depending on whether information on other attributes A_2 or A_3 is provided.

Imagine a world in which there are only three relevant features, A_1, A_2, A_3 that combine to generate evaluation Y of a candidate according to the law:

$$Y = A_2 \times A_3.$$

So $Y = 1$ if and only if $A_2 = 1$ and $A_3 = 1$; A_1 plays no role in this law.

Let us ask: would a subject prefer a candidate with $A_1 = 1$ over a candidate with $A_1 = 0$? Or, equivalently: does A_1 affect preferences for a candidate?

The answer might appear to be "no." Certainly, if a voter were informed about A_2 and A_3 they could figure out Y and their assessment would not depend on A_1. So in a full information world the answer is no.

But what if they did not know A_2 or A_3? Well then, A_1 could matter, even though it is not part of the data generating process for Y, because it could be informative for A_2 and A_3, which are.

To illustrate, here is an example in which seeing $A_1 = 1$ increases beliefs about Y by 0.25 compared to the case when $A_1 = 0$, when neither A_2 or A_3 is observed. But seeing $A_1 = 1$ *decreases* beliefs about Y by 0.25 compared to the case when $A_1 = 0$, when neither $A_2 = 1$ is observed but A_3 remains observed. The takeaway of course is that these are all correct: whether A_1 weighs positively or negatively or not at all in preferences depends entirely on what other information is available.

For the illustration imagine the following joint distribution over A_1, A_2 and A_3:

	$A_1 = 0$	$A_1 = 1$
$A_2 = 0,\ A_3 = 0$	$\frac{32}{16}$	$\frac{8}{16}$
$A_2 = 0,\ A_3 = 1$	$\frac{27}{16}$	$\frac{9}{16}$
$A_2 = 1,\ A_3 = 0$	$\frac{2}{16}$	$\frac{18}{16}$
$A_2 = 1,\ A_3 = 1$	$\frac{6}{16}$	$\frac{18}{16}$
Total	$\frac{12}{16}$	$\frac{4}{16}$

Then, recalling that $\Pr(Y = 1) = \Pr(A_2 = 1 \& A_3 = 1)$ it is easy to check that:

1. $\Pr(Y = 1|A_1 = 1) - \Pr(Y = 1|A_1 = 0) = \frac{1}{2} - \frac{1}{4} = 0.25$
2. $\Pr(Y = 1|A_1 = 1 \& A_2 = 1) - \Pr(Y = 1|A_1 = 0 \& A_2 = 1) = \frac{1}{2} - \frac{3}{4} = -0.25$
3. $\Pr(Y = 1|A_1 = 1 \& A_2 = 1 \& A_3 = 1) - \Pr(Y = 1|A_1 = 0 \& A_2 = 1 \& A_3 = 1) = 0.$

The example makes clear that unless we are accessing some form of full information benchmark, the estimand—"effect" of features on preferences—can be zero, positive, or negative. It depends on what other information is available to individuals, and of course subjects' beliefs about the informativeness of one signal about another. By the same token *pooled* estimates of effects across studies will depend critically on the distribution of which factors are controlled across different studies.

D Does control reduce uncertainty?

Commonly we think that one motivation for controlling third factors (both an the intervention and analysis stages) is to ensure tighter estimates of treatment effects. This is not always the case however and so in survey experiments controlled variation in third factors may mean more rather than less variation.

To illustrate, say in truth for all individuals $Y = A_1 A_2$ where A_1 and A_2 are beliefs about features of candidates and Y is an evaluation. In the target application, A_2 has a known distribution, say 50% of the population of candidates is male and this is known to all individuals. We will assume that for any candidate profile subjects will always form beliefs about the features and that we are interested in the effects of these beliefs. Moreover beliefs about candidate profiles are fully determined by candidate descriptions if these exist. (These provisions are to meaningfully keep the estimand constant when we introduce new information about a candidate.)

Given information about A_1 only, imagine all individuals form evaluations in line with expected utility: $\hat{Y} = \mathbb{E}[Y] = 0.5 \times A_1 + 0.5 \times 0$. Varying A_1 alone will allow perfect estimates of the average effect of beliefs A_1: 0.5 (individuals observing $A_1 = 1$ will report 0.5, individuals observing $A_1 = 0$ will report 0). There is no error. Say now that we vary A_2, presenting 0s and 1s each to half the subjects in a balanced manner. Then individuals observing $A_1 = 1$ will report 0 or 1 depending on A_2. In that case if we regress Y on A_1 we will get the same answer but with less precision. If we include a linear control for A_2 we do better, but still worse than if we had not introduced variation in A_2 at all. If we use a saturated model (e.g. following Lin (2013)), we get back to where we started: effectively undoing at the analysis stage the noise that we generated by adding the control feature at the intervention stage.

The key point is simply that in this setting, controlling features at the intervention stage does not necessarily reduce variance, it can increase it. In a typical field experiment controlling a background feature is motivated by the idea that the background feature would take on natural values and vary in any case. But in a factorial experiment the candidates do not exist outside of the experiment; at best we have the beliefs that individuals have and that they project onto these candidates. And these beliefs may not exhibit variation even if they are formed with respect to a supposed background distribution of features that does vary. Thus controlled variation can induce variation in beliefs where none existed before.

This example hinges of course on the idea that at baseline, absent information on A_2, there is little variation in beliefs about A_2. If instead there was variation in beliefs then controlling beliefs and implementing a saturated regression would remove noise.

In that case however the interpretation of the estimand is a little more complicated. If we are interested, for instance, in the effect of providing information about A_1, absent information on A_2, then this effect will vary as a function of expectations of A_2, which we now believe also vary. Moreover an inference from the experiment where we control beliefs to effects for strata with different beliefs (e.g. .5) requires extrapolation of the form we did above when we assumed individuals act on expectations following the expected utility model.

E Changing estimands or identification failures?

I describe how the set of controls changes the estimand values. Dafoe, Zhang, and Caughey (2018) engages with a similar logic but concludes rather that beliefs are not identified in these models.

They write: "Only if the IE assumption holds can response differences between versions of the survey be attributed to differences in subjects' beliefs about the factor of interest." Yet here the assumption is obviously violated, but no bias is introduced: rather the effects of an intervention of beliefs operates *via* inferences on beliefs about other attributes, even if these are not downstream from factors of interest in the subjects causal model. In fact the causal relations between attributes have no part at all to play here. X might cause Y yet being informed about Y alters beliefs about X. They are interested implicitly in a kind of controlled effect, the effect of change in beliefs about D *without allowing the change in beliefs about B which a change in beliefs about D entails.*

Concretely, in World 2, in the account of Dafoe, Zhang, and Caughey (2018), we should be assessing the effects of beliefs about W and P while keeping beliefs about A fixed. The estimand would then be 0 in World 2 in all cases in which information about A is not provided. The interpretation here is different as it takes the adjustment in beliefs about A to be downstream: beliefs about A here are certainly downstream with respect to the treatment – the signal about W for instance. It is admittedly less clear whether we can think of the updating on secondary attributes as being downstream with respect to the updating on treated attributes.

The downstream interpretation works if we think that subjects first update on the feature for which they received information, and then update on other attributes—for instance you learn someone is from a given identity group and you then make inferences about their skills. It is not so clear if the updating is conceived as simultaneous across factors. So the question comes down to a somewhat fine point of whether belief updating on a second factor can be thought of as a consequence of direct manipulation of beliefs about another factor, or as a consequence of the same intervention.

Table 14: ATE comparisons for A_1 in World 1

Rival	p	Comparison	Diff
(0,0)	0.4	$.6(\Pr(U(1,1) > U(0,0)) - \Pr(U(0,0) > U(0,0)))+$	.3
		$.2(\Pr(U(1,1) > U(0,0)) - \Pr(U(0,1) > U(0,0)))+$	
		$.2(\Pr(U(1,0) > U(0,0)) - \Pr(U(0,0) > U(0,0)))$	
(0,1)	0.1	$.6(\Pr(U(1,1) > U(0,1)) - \Pr(U(0,0) > U(0,1)))+$	.3
		$.2(\Pr(U(1,1) > U(0,1)) - \Pr(U(0,1) > U(0,1)))+$	
		$.2(\Pr(U(1,0) > U(0,1)) - \Pr(U(0,0) > U(0,1)))$	
(1,0)	0.1	$.6(\Pr(U(1,1) > U(1,0)) - \Pr(U(0,0) > U(1,0)))+$	.3
		$.2(\Pr(U(1,1) > U(1,0)) - \Pr(U(0,1) > U(1,0)))+$	
		$.2(\Pr(U(1,0) > U(1,0)) - \Pr(U(0,0) > U(1,0)))$	
(1,1)	0.4	$.6(\Pr(U(1,1) > U(1,1)) - \Pr(U(0,0) > U(1,1)))+$	.3
		$.2(\Pr(U(1,1) > U(1,1)) - \Pr(U(0,1) > U(1,1)))+$	
		$.2(\Pr(U(1,0) > U(1,1)) - \Pr(U(0,0) > U(1,1)))$	

F Calculations for illustration (Section 4)

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis* 26 (4): 357–78.

Amsalem, Eran, and Alon Zoizner. 2024. "The Causal Effect of Candidate Extremity on Citizens' Preferences: Evidence from Conjoint Experiments." *Public Opinion Quarterly* 88 (3): 859–85.

Anderson, Norman H. 1971. "Integration Theory and Attitude Change." *Psychological Review* 78 (3): 171.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. 2023. "Using Conjoint Experiments to Analyze Election Outcomes: The Essential Role of the Average Marginal Component Effect." *Political Analysis* 31 (4): 500–518. https://doi.org/10.1017/pan.2022.16.

Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104 (2): 226–42.

Bell, Mark S, and Kai Quek. 2018. "Authoritarian Public Opinion and the Democratic Peace." *International Organization* 72 (1): 227–42.

Blair, Graeme, Alexander Coppock, and Macartan Humphreys. 2023. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign*. Princeton University Press.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20 (1): 47–77.

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. "Design and Analysis of the Randomized Response Technique." *Journal of the American Statistical Association* 110 (511): 1304–19.

Boittin, Margaret L, Rachel S Fisher, and Cecilia Hyunjung Mo. 2024. "Evidence of Caste-Class Discrimination from a Conjoint Analysis of Law Enforcement Officers." *American Political Science Review* 118 (1): 504–11.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26 (4): 399–416.

De la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. 2022. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis* 30 (1): 19–45. https://doi.org/10.1017/pan.2020.40.

Druckman, James N. 2001. "The Implications of Framing Effects for Citizen Competence." *Political Behavior* 23 (3): 225–56.

Fang, Al, Andrew Guess, and Macartan Humphreys. 2018. "Can the Government Deter Discrimination." *The Journal of Politics*.

Fisher, Ronald Aylmer. 1971. *The Design of Experiments*. Springer. https://archive.org/details/in.ernet.dli.2015.502684/mode/1up?q=factorial.

Ganter, Flavien. 2023. "Identification of Preferences in Forced-Choice Conjoint Experiments: Reassessing the Quantity of Interest." *Political Analysis* 31 (1): 98–112.

Goldin, Jacob, and Daniel Reck. 2015. "Framing Effects in Survey Research: Consistency-Adjusted Estimators." *Unpublished, Stanford Law School, CA, US*.

———. 2020. "Revealed-Preference Analysis with Framing Effects." *Journal of Political Economy* 128 (7): 2759–95.

Green, Donald P, Shang E Ha, and John G Bullock. 2010. "Enough Already about 'Black Box' Experiments: Studying Mediation Is More Difficult Than Most Scholars Suppose." *The Annals of the American Academy of Political and Social Science* 628 (1): 200–208.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112 (8): 2395–2400.

Hartmann, Felix, Macartan Humphreys, Ferdinand Geissler, Heike Klüver, and Johannes Giesecke. 2024. "Trading Liberties: Estimating Covid-19 Policy Preferences from Conjoint Data." *Political Analysis* 32 (2): 285–93. https://doi.org/10.1017/pan.2023.25.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Humphreys, Macartan, and Alexandra Scacco. 2020. "The Aggregation Challenge." *World Development* 127: 104806.

Kertzer, Joshua D, Jonathan Renshon, and Keren Yarhi-Milo. 2021. "How Do Observers Assess Resolve?" *British Journal of Political Science* 51 (1): 308–30.

Kirkland, Patricia A, and Alexander Coppock. 2018. "Candidate Choice Without Party Labels: New Insights from Conjoint Survey Experiments." *Political Behavior* 40 (3): 571–91.

Kuhn, Patrick M, and Nick Vivyan. 2022. "The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries." *Political Analysis* 30 (3): 381–402.

Landgrave, Michelangelo, and Nicholas Weller. 2022. "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment." *Political Analysis* 30 (1): 142–48.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics*, 295–318.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–65.

Mares, Isabela, and Giancarlo Visconti. 2020. "Voting for the Lesser Evil: Evidence from a Conjoint Experiment in Romania." *Political Science Research and Methods* 8 (2): 315–28.

McKenzie, Craig RM, Shlomi Sher, Xingyu Liu, and Leo J Kleiman-Lynch. 2025. "When and Why Framing Effects Are Neither Errors nor Mistakes." *Mind & Society*, 1–21.

Olinger, Reilly, Benjamin Matejka, Rohan Chakravarty, Margaret Johnston, Eliana Ornelas, Julia Draves, Nishi Jain, et al. 2024. "Americans Do Not Select Their Doctors Based on Race." *Frontiers in Sociology* 8: 1191080.

Ono, Yoshikuni, and Barry C Burden. 2019. "The Contingent Effects of Candidate Sex on Voter Choice." *Political Behavior* 41 (3): 583–607.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100 (469): 322–31.

Samii, Cyrus. 2025. "Survey Experiments and the Credibility Revolution." 2025. https://cyrussamii.com/?p=4168.

Schachter, Ariela, and Katherine Weisshaar. 2025. "Survey Experiments in Sociology." *Annual Review of Sociology* 51 (1): 149–69.

Schwarz, Susanne, and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84 (2): 655–68.

Slovic, Paul. 1995. "The Construction of Preference." *American Psychologist* 50 (5): 364.

Sniderman, Paul M, and Thomas Piazza. 1993. *The Scar of Race*. Harvard University Press.

Tamer, Elie. 2010. "Partial Identification in Econometrics." *Annu. Rev. Econ.* 2 (1): 167–95.

Tipton, Elizabeth. 2021. "Beyond Generalization of the ATE: Designing Randomized Trials to Understand Treatment Effect Heterogeneity." *Journal of the Royal Statistical Society Series A: Statistics in Society* 184 (2): 504–21.

Tomz, Michael R, and Jessica LP Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107 (4): 849–65.

Torreblanca, Carolina, William Dinneen, Guy Grossman, and Yiqing Xu. 2025. "The Credibility Revolution in Political Science."

Tversky, Amos, and Daniel Kahneman. 1987. "Rational Choice and the Framing of Decisions."

VanderWeele, Tyler J. 2018. "On Well-Defined Hypothetical Interventions in the Potential Outcomes Framework." *Epidemiology* 29 (4): e24–25.

VanderWeele, Tyler J, and Miguel A Hernan. 2013. "Causal Inference Under Multiple Versions of Treatment." *Journal of Causal Inference* 1 (1): 1–20.

Von Winterfeldt, Detlof, and Gregory W Fischer. 1975. "Multi-Attribute Utility Theory: Models and Assessment Procedures." In *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973*, 47–85. Springer.

Ylikoski, Petri. 2013. "Causal and Constitutive Explanation Compared." *Erkenntnis* 78 (Suppl 2): 277–97.