Three confusions in the use and interpretation of survey 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 concommitant confusion regarding what the goal of a survey experiment is. Focussing especially on conjoint survey experiments, I highlight three points of confusion and provide pointers to clarify the targets of inference. 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 licence 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. In the case of conjoint experiments, for instance, there is interest in understanding the effects of candidate *attributes* on vote shares, for isntance, rather than simply the effects of *signals* (within a bundle of signals) on reported preferences or hypothetical choices. 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 short note 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 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 substanitively altering the estimand. This is done in multiple ways. Often researcher 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 exogeneous—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" down stream 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 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 is 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 the figure below, preferences and features combine to form choices. By altering features and observing choices we learn about preferences.

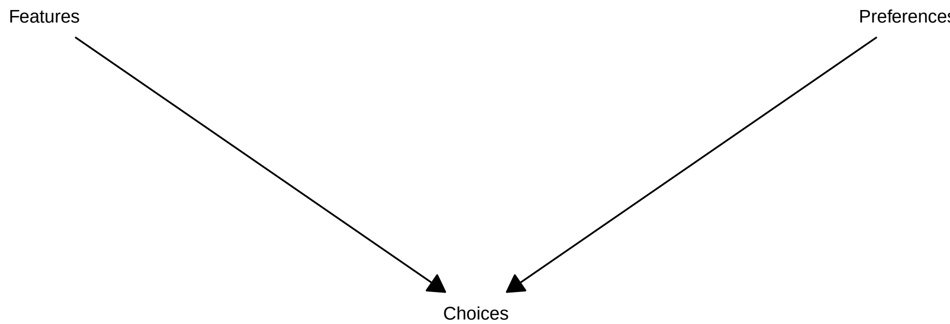


Figure 1: A model in which preferences and features jointly determine choices. Learning about the effect of features on choices lets you make inferences about preferences.

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).

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 say it's because it got hit by a bus, you are giving a causal account to explain the state it's in. If you say: because turning the peddles does not make the bike go forward, you are giving a causal account that redescribes the problem. If you say: because the wheels are missing, you are giving an account that depends on a property of the bike. You are telling me about the "causal capacity" of the bike.

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 for **descriptive inference**, consider the **randomized response** survey experiments. In randomized response experiment, people are randomly assigned to answer either a sensitive question or a non-sensitive question 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 sections unpack the ideas in this table.

2.2.1 Priming experiments

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

2.2.1.1 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.

2.2.1.2 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.

2.3 List experiments

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

2.3.1 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.

2.3.2 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.

2.4 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).

2.4.1 For descriptive inference

Goldin and Reck (2015) (see also Goldin and Reck (2020)) treat framing effects as more of a nuisance than a quantitiy 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 pont identify quantities of interests specifically for subgroups who are unaffected by frames.

2.4.2 For causal inference

Many studies examining framing effects are specifically interested in causal quntities, specifically the effects of various types of triggers on the ways people think about issues. The basic example of a loss verseus 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.[2]

2.5 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

[2]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 expressionof preferences, are affected.

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).

2.5.1 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.

2.5.2 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 features 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. This may of course 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.[3]

The risk above remains, however: the effect you are getting is the effect of the attribute 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

[3]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.

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).

[[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 irrelavant to the choice, given the options; be prominment in preferences. I address this point from a different angle later, suggesting that "preferences" in these context are often best conceived as menu dependent.]]

3 The role of controls

Including controls in an experiment can be done for multiple purposes.

To clarify terminology, we can think of controlling at the intervention stage: fixing experimental conditions or at the analysis stage, taking account of the feature 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. In survey experiments researchers can similarly introduce controls at both stages – as part of their data strategy and part of their answer strategy. Note also that at the analysis stage researchers often use the term "controlling for" when referring to the introduction of "linear controls," for instance in a regression. More strictly however the term is used for the idea of completely conditioning on the factor – examing effects at all possible values of the factor (given the assignment) and then determining an average effect.

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.[4]

[4] Fisher highlighting 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.

In addition there are at least three arguments for using controls at the analysis stage:

1. Remove bias, or confounding
2. Reduce variance
3. Target specific quantities of interest

In a survey experiment introducing these controls at the analysis stage presupposes their introduction at the intervention stage (that is, for "candidate level" controls; subject level controls and similar can of course also be used).

3.1 Addressing bias

Controls in survey experiments are sometimes described at helping address confounding. It is a somewhat curious claim since random assignment itself ought to remove confounding.

Users worry about confounding however. Bell and Quek (2018) write for instance 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." Tomz and Weeks (2013) also describe their concern about leaving out factors "that could confound the relationship between shared democracy and public support for war."

Dafoe, Zhang, and Caughey (2018) write "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)" and "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." But it surely does alter beliefs about the focal attribute which in turn operates through beliefs about other attributes.

If one is interested just in the effects of receiving information, given whatever you believe already, then the effect is identified just from random assignment. 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 recieve 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 we are not talking about the problem that in thre world the effect of democracy on alliances is confounded by trading relations.

Dafoe, Zhang, and Caughey (2018)'s focus however is not on the effects of information about a factor, but rather about the effects of beliefs about the factor, which themselves may be manipulated via manipulation of information, uncontaminated by effects working through beliefs of other factors.

The dependence of the interpretation of average effects 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. Although one might have a notion of fundamental preferences—what would be valued in a full information environment—is settings with imperfect information, preferences over options depedn critically on beliefs of unknown features of those options.

3.2 Reduce uncertainty?

Commonly we think that one motivation for controlling third factors is to ensure tighter estimates of treatment effects. THis is now always compelling in the case of survey experiments where controlled variation in third factors may mean more rataher than less variation.

Say in truth for all individuals $Y = X_1X_2$ where $X1$ and X_2 are beliefs about features of candidates and Y is anevaluation. In the target application, X_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 meanigfully keep the estimand constant when we introduce new information about a candidate.)

Given information about X_1 only, imagine all individuals form evaluations in line with expected utility: $\hat{Y} = \mathbb{E}[Y] = 0.5 \times X_1 + 0.5 \times 0$. Varying X_1 alone will allow perfect estimates of the average effect of beliefs X_1: 0.5 (individuals observing $X_1 = 1$ will report 0.5, individuals observing $X_1 = 0$ will report 0). There is no error. Say now that we vary X_2, presenting 0s and 1s each to half the subjects in a balanced manner. Then individuals observing $X_1 = 1$ will report 0 or 1 depending on X_2. In that case if we regress Y on X_1 on we will get the same answer but with less precision. If we include a linear control for X_2 we do better, but still worse than if we had not introduced variation in X_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 X_2, there is little variation in beliefs about X_2. If instead there was variation in beliefs then controlling beliefs and implementing a starutated regression would remove noise.

In that case however the interpretation of the estimand is a little more complicated. If we are interested, for isntance, in the effect of providing information about X_21, absent information on X_2, then this effect will vary as a function of expectations of X_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.

3.3 Control dependent beliefs

Imagine a world in which there are only three relevant features, X_1, X_2, X_3 that combine to generate evaluation Y of a candidate according to the law:

$$Y = X2 \times X3.$$

So $Y = 1$ if and only if $X_2 = 1$ and $X_3 = 1$; X_1 plays no rule in this law.

Let us ask: would a subject prefer a candidate wit $X_1 = 1$ over a candidate with $X_1 = 0$? Or, equivalently: does X_1 affect preferences for a candidate?

The answer might appear to be "no." Certainly, if a voter were informed about X_2 and X_3 they could figure out Y and their assessment would not depend on X_1. So in a full information world the answer is no.

But what if they did not know X_2 or X_3? Well then, X_1 could matter, even though it is not part of the data generating process for Y, because it could be informative for X_2 and X_3, which are.

To illustrate, here is an example in which seeing $X_1 = 1$ increases beliefs about Y by 0.25 compared to the case when $X_1 = 0$, when neither X_2 or X_3 is observed. But seeing $X_1 = 1$ *decreases* beliefs about Y by 0.25 compared to the case when $X_1 = 0$, when neither $X_2 = 1$ is observed but X_3 remains observed. The takeaway of course is that these are all correct: whether X_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 X_1, X_2 and X_3:

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0,\ X_3 = 0$	$\frac{32}{16}$	$\frac{8}{16}$
$X_2 = 0,\ X_3 = 1$	$\frac{27}{16}$	$\frac{9}{16}$
$X_2 = 1,\ X_3 = 0$	$\frac{2}{16}$	$\frac{18}{16}$

	$X_1 = 0$	$X_1 = 1$
$X_2 = 1,\ X_3 = 1$	$\frac{6}{16}$	$\frac{18}{16}$
Total	$\frac{12}{16}$	$\frac{4}{16}$

Then, recalling that $\Pr(Y = 1) = \Pr(X_2 = 1 \& X_3 = 1)$ it is easy to check that:

1. $\Pr(Y = 1|X_1 = 1) - \Pr(Y = 1|X_1 = 0) = \frac{1}{2} - \frac{1}{4} = 0.25$
2. $\Pr(Y = 1|X_1 = 1 \& X_2 = 1) - \Pr(Y = 1|X_1 = 0 \& X_2 = 1) = \frac{1}{2} - \frac{3}{4} = -0.25$
3. $\Pr(Y = 1|X_1 = 1 \& X_2 = 1 \& X_3 = 1) - \Pr(Y = 1|X_1 = 0 \& X_2 = 1 \& X_3 = 1) = 0$.

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 dsitribution of which factors are controlled across different studies.

A key 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.

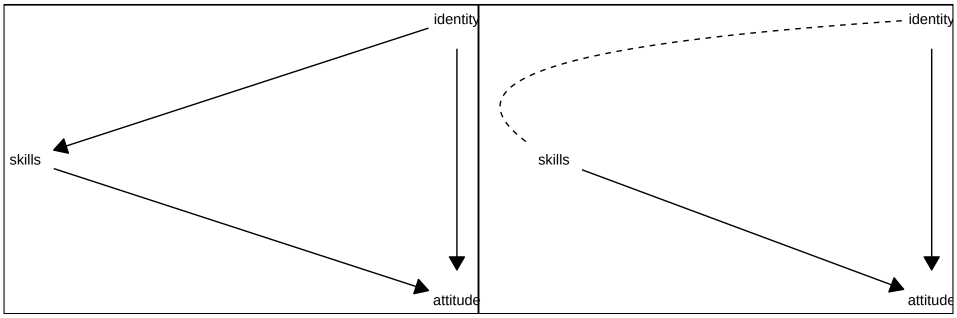
3.4 Controlled direct effects and taste-based discrimination

A particular usage of controls to identify a discriminatory behavior, in particular to assess how judgements depend on one attribute – race, gender, for instance—while specifically controlling for other features that are plausibly associated with teh attribute, at least in the minds of subjects.

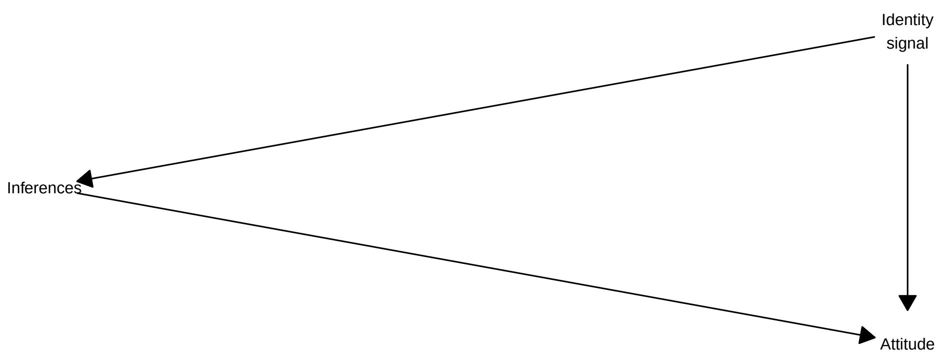
Boittin, Fisher, and Mo (2024) suggest that discrimination *after* controlling for criminality suggests taste based discrimination (though they consider other possibilities also). Ono and Burden (2019) argue that controlling for other features lets them assess taste based discrimination. Olinger et al. (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.

(**fig:idg?**)



In the real world causal process the attributes may or may not be post treatment with respect to identity, but in the beliefs model they are post treatment (model 3). However if they are exogeneously controlled by the experimenter, we can estimate the controlled direct effect of the identity cue.



Two subtleties:

- Controlling may produce a controlled direct effect, is this a a more natural quantity than the natural direct effect? Yes, in particular sense: it does capture preferences but not behavior in the following sense. The natural direct effect captures a situation where there are beliefs about what skills they would they expect if they thought the candidate were a man; upon learning the candidate were a woman, how would they evaluate teh candidate if they thought the candidate had the skills they associated with a man? Say now someone has the follow proclivities. They value someone 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 makes a difference—regardless of the actual distribution of skills in the world. THe natural direct effect of gender is 0. However the controlled direct effect of gender is 1 conditional on low skill and 0 otherwise.

Controlling you are conditioning on uniform, as highlighted by others, but would need to base it on the distribution of *beliefs* — not the distribution of attributes.

- The meaning of the estimand depends crucially on what attributes get controlled for. Indeed if we imagine that "taste" effects work thorugh particular channels: disgust, affection, beliefs about integrity or piousness — then controlling for these effects woudl remove direct effects entirely. A Heisenberg principle. We cannot both emasure and understand direct effects.

General claims such as "Americans do not select their doctors based on race" Olinger et al. (2024) are misleading if one has controlled away various effects of the attribute (or of the signal of the attribute).

3.5 Mental models

General idea that including more factors increases control; see for instance the discussion of "full profile" strategies in Green and Srinivasan (1978).

The key idea here is that although a feature might be randomized in the presentation of a description of a candidate, that does not mean that the feature is exogeneous with respects to other attributes that respondents project onto imagined candidates. For example if I tell you that a candidate is criminal you might infer that they are male. This inference might blocked if I also report the candidate's gender, but inferences to features that are not included are not blocked. In that sense randomization of signals about candidates does not do the same work as randomization of attributes themselves. In this setting, following the logic of "M-bias"—that can arise for treatments that are not randomily assigned— it is even possible that biases enter *because* you control for additional features.

3.5.1 M-bias illustration

I illustrate using an example of "conditioning on a collider." [Pearl]

Consider a conjoint experiment in which respondents evaluate hypothetical political candidates. Each candidate profile randomly assigns the following attributes:

- W: Wealth (wealthy vs. not wealthy), treated
- P: Privilege, whether the individual comes from an elite background

and voters report on Y, beliefs about politician quality.

Say we are interested in how P affects Y, and include W as a background condition.

Both W and P are fully randomized by the experimenter.

There is a causal effect on beliefs: changing the information about privilege changes the beliefs about Y, given the mental model. But *in* the mental model, privilege has no effect on quality.

3.5.2 Respondents' mental model

Although attributes are randomized in the presentation of profiles, *respondents do not believe they are random* for candidates in their modental models of the world. Instead, they reason causally about how politicians acquire wealth. In particular, respondents posit a latent trait, ability, that can combine with privilege to produce wealth.

Represented as a DAG their mental causal model is:

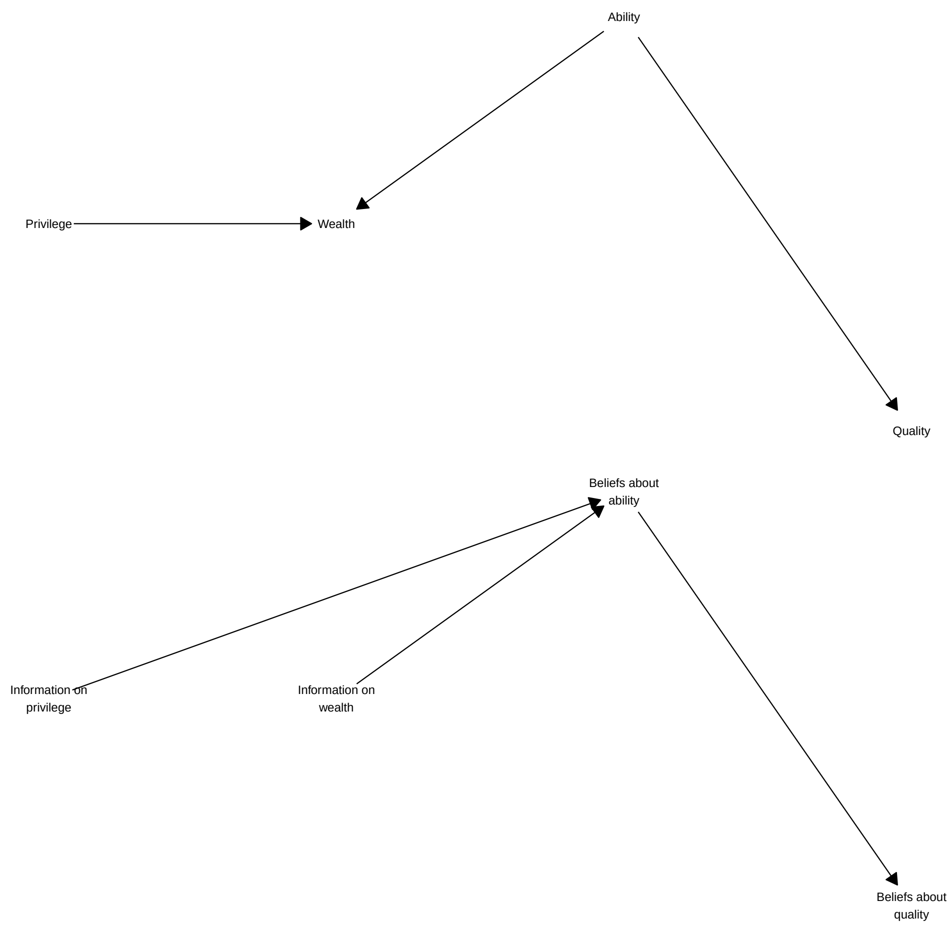


Figure 2: Collider bias can render a *signal* of a feature relevant to assessments even though– in the individual's mental model of the world—the feature itself does not affect assessments.

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

We can see citizens do not think that wealth causes quality, but it is absolutely *informative* about quality, since it suggests ability. Privilge however is not (absent information on welath) informative about quality at all: it neither has an effect nor carries informaiton.

Privilege and ability are distributed independently. Ability predicts wealth prefectly. Either privilege or ability can give rise to wealth, but they are jointly sufficient for wealth. If each of these four data types are equally likely we have the following inferences about Y:

$$\Pr(Y=1|P=1,W=1)=\frac{2}{3}<1=\Pr(Y=1|P=0,W=1)$$

$$\Pr(Y=1|P=1,W=0)=0<\frac{1}{3}=\Pr(Y=1|P=0,W=0)$$

Conditional on W–regardless of what W is, seeing privilege decreases your belief in quality by $\frac{1}{3}$.

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.

4 Mapping from estimates from survey experiments to real world estimands

4.1 The prime mapping

4.2 The conjoint mapping

Researchers often appear to draw quite strong conclusions about real world processes on the basis of findings from conjoint 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]

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

This is an interesting example 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 multidude 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 linkages problems since there is a well defined aggregation rule that one should be able to plug in to make the link.

In the remaining discussion I suggest that even under optimistic conditions this might be hard. I point to multiple hurdles. One of these is particularly important however: when there are multiple attributes, an exogenous change in one may result in an endogenous change in

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

another. Conjoints are used to deliver a kind of controlled effect—controlling not just for confounds but also for the causal pathways along which treatment effects might naturally operate. Controlled effects may not depend on causal relations between attributes, but the effect of a single treatment, as usually understood, absolutely can.[6]

4.3 Licence to export

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 os 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 infomration 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 of vote shares given the actual candidates running in a race, but the expected effect across races that might have been run).[7]

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).

4.3.1 Structure

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: $v_j(a)$.

[6]To avoid confusion, the issue is not about the joint distribution of attributes, but about the *underlying causal structure* relating attributes to one another and to voter utilities.

[7]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.

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.

4.3.2 The ATE

Let $\beta(u)$ 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 g 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. Thus the meaning of β depends on h in an obvious way.

4.3.3 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.

Write individual-level potential outcomes as:

$$y_i(s^i) \in \{0, 1\}.$$

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

$$\tau_i \equiv y_i(1, s^i_{-1}) - y_i(0, s^i_{-1}),$$

where s^i_{-1} 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

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

Finally, if we are interested in the expectation of this treatment effect under a distribution f over background information profiles s_{-1}, we define

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

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.

As an aside note that τ_i written as a controlled effect. 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, this is equivalent to the familiar unit level treatment effect $y_i(1) - y_i(0)$. Again under the exclusion restriction we can suppress the background information and simply write the familiar quantity $\tau = \mathbb{E}[Y(1) - Y(0)]$, where we use $Y(.) \equiv y_i(.)$ with $i \sim \mu$ to denote the random potential outcome induced from sampling i from N.

4.3.4 Equivalence

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

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

I give an answer using the following conditions which collectively are sufficient for the equivalence claim.

1. **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 attribute g does not causally affect other attributes. Attributes may of course be correlated with each other.

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

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

   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 axiom may be considered a "no misreporting" condition.

3. **Rational but nonstrategic voting.**
   Let $y_i^j(a)$ indicate whether candidate j is i's preferred candidate given attribute vector a, and let $v_i^j(a)$ indicate whether i votes for j.

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

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.

4. **Context irrelevance given attributes.**
   For all contexts u and attribute vectors a,

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

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

5. **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

$$y_i = \begin{cases} 1 & \text{if } u_i^j(a) \geq u_i^k(a), \\ 0 & \text{otherwise.} \end{cases}$$

6. **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 1–6 hold, then the expected effect of a change in an attribute over contexts on vote shares of candidate j (β) equals the expected effect of a change in information about that attribute on expressed preferences, conditional on other information (AMCE).

We have a direct proof in six steps:

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

4.4 Easy and hard cases

For a positive answer consider a causal model like that shown in Figure 3.

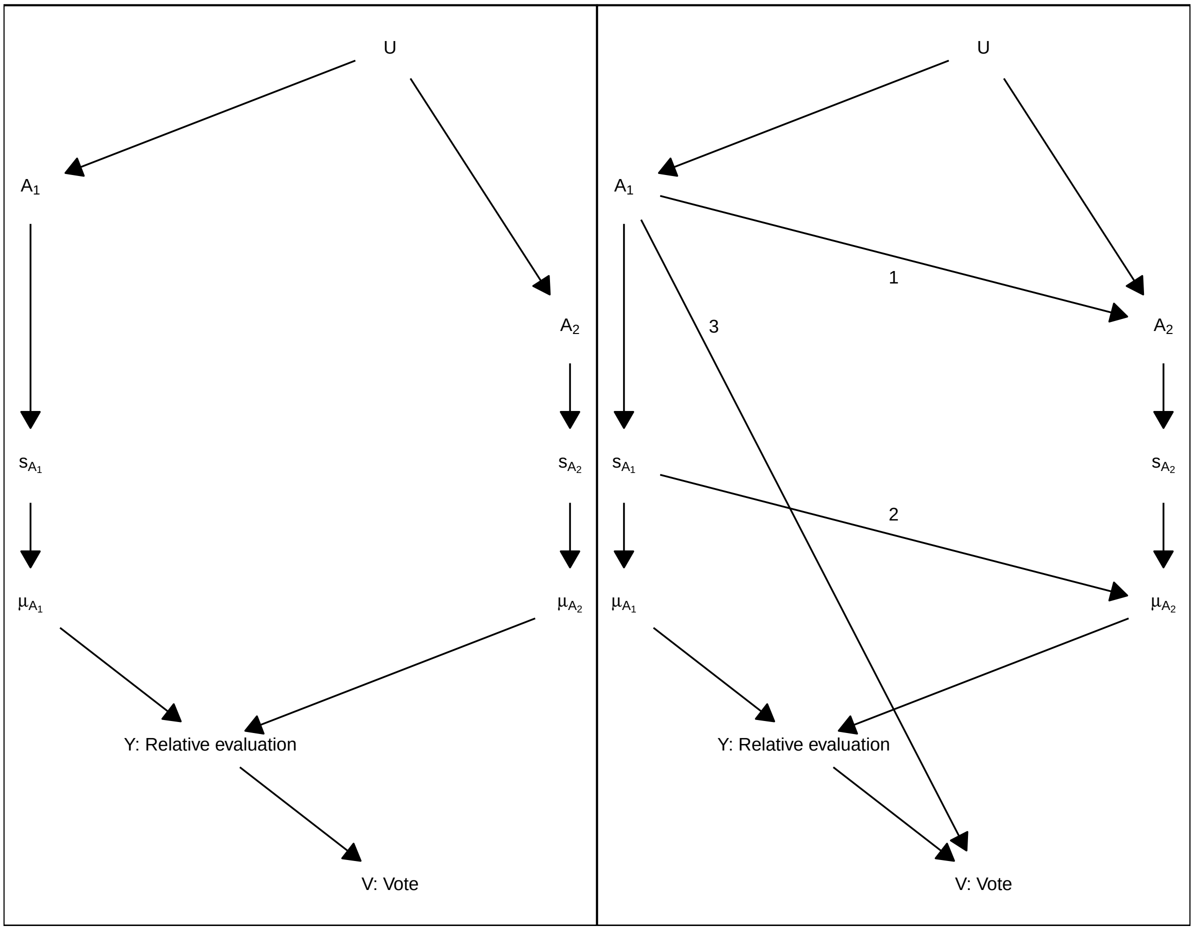


Figure 3: DAG showing relation between two features, A_1 and A_2, an unobservable feature U, two *signals* about these observable features s_1 and s_2, respondent *beliefs* about these features μ_1 and μ_2, and a respondent level (relative) evaluation, Y, and a vote choice. A survey experimentalist intervenes 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.

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 woudl 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.

4.4.1 Mutually endogeneous 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.

Rather we would need to set our sights on some form of controlled direct effect of A_1 on Y, blocking the effects that run through A_2. How meaningful this estimand is surely depends on the case. And it can depend sharply on the set of attributes that happen to be included in the experiment. It is not hard to imagine, moreover, that the controlled effect could ultimately be thin soup. In a line of dominos, each have a unit effect on the fall of the final domino, only the second last domino has a non zero controlled direct effect. More sharply: one could make any controlled direct effect "vanish" by inclusion of features that lie on all pathways from the node to the outcome.

To me 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.

4.4.2 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)).

4.4.3 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.

4.5 Numerical example

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 XX. 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%, $A_1 = 2$ regardless of A_1. 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 A and B (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 A, holding A_2 fixed, across all possible profiles of candidate B. 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 B as given, a change in A1 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. This is less obvious to see byt the Table below lays out the calculations explicitly.

Table A2: ATE comparisons for A_1 in World 1

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

In World 2, changing A_1 does not affect A_2. Table A3 reports the resulting comparisons.

$$\text{ATE}_1^{\text{World 2}} = 0.$$

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.

5 Conclusion

The issue here is not unique to conjoints of course. It arises also in *audit experiments*—such as that studied in Fang, Guess, and Humphreys (2018), in which myriad features of fictional individuals are assigned independently. Doing so makes sense if our interest is indeed on different types of discrimination rather than the effects of "being" from one group or another. Since for this we may be specifically interested in the effect of a feature — such as race, on responses — conditional not just on possible confounders but on post-treatment features also.

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, Paul E, and Venkatachary Srinivasan. 1978. "Conjoint Analysis in Consumer Research: Issues and Outlook." *Journal of Consumer Research* 5 (2): 103–23.

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.

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."

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.