NBER WORKING PAPER SERIES

CAUSAL INFERENCE FROM HYPOTHETICAL EVALUATIONS

B. Douglas Bernheim
Daniel Björkegren
Jeffrey Naecker
Michael Pollmann

Working Paper 29616
http://www.nber.org/papers/w29616

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2021, Revised February 2026

Thank you to seminar and conference participants for helpful comments. Detailed suggestions from Richard Carson and Laura Taylor were especially helpful. This paper is related to a previous working paper, "Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions," by Bernheim, Björkegren, Naecker, and Antonio Rangel; it uses data from the same lab experiment, but most of the methodological analysis is new, as is the field application. We are grateful to Antonio Rangel for his contributions to the earlier project and to Irina Weisbrott for assistance with collecting the laboratory data. Bernheim acknowledges financial support from the National Science Foundation through grant SES-1156263. Björkegren thanks the W. Glenn Campbell and Rita Ricardo-Campbell National Fellowship at Stanford University, and Microsoft Research for support. Pollmann was supported generously by the B.F. Haley and E.S. Shaw Fellowship for Economics through a grant to the Stanford Institute for Economic Policy Research. The components of this study were overseen by the IRBs of Stanford University or Brown University. The field experiment in this study was pre-registered with the AEA RCT Registry (AEARCTR-0004885); the lab experiment was conducted prior to the establishment of the registry. An accompanying R package is available on Github: https://github.com/michaelpollmann/hypeRest. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Causal Inference from Hypothetical Evaluations
B. Douglas Bernheim, Daniel Björkegren, Jeffrey Naecker, and Michael Pollmann
NBER Working Paper No. 29616
December 2021, Revised February 2026
JEL No. C13, D12

ABSTRACT

We develop methods for estimating the effects of treatments on choices by combining data on real choices with hypothetical responses. Our basic method accounts for potential variation of hypothetical bias across treatment states yet is applicable even when treatment variation in real choice data is absent or limited. Under stated conditions, it also yields consistent estimates in applications with endogenous treatments. For applications that do not satisfy those conditions, we offer variants of the approach that do not require the analyst to identify sources of quasi-experimental variation. We provide formal econometric theory and test the methods in two applications.

B. Douglas Bernheim
Stanford University
Department of Economics
and NBER
bernheim@stanford.edu

Daniel Björkegren
Columbia University
dan@bjorkegren.com

Jeffrey Naecker
Google
jnaecker@google.com

Michael Pollmann
Duke University
michael.pollmann@duke.edu

A data appendix is available at http://www.nber.org/data-appendix/w29616

1 Introduction

It is often challenging to estimate how choices respond to public policies and private interventions. A potential solution is to rely on what people *say* they would choose. One strand of literature treats hypothetical choices as legitimate measures of real choices (Krueger and Kuziemko 2013; Kuziemko et al. 2015), at least when particular protocols are followed (e.g., Cummings and Taylor 1999; Jacquemet et al. 2013, Blamey, Bennett, and Morrison 1999; Loomis, Traynor, and Brown 1999; and Mas and Pallais 2017 and 2019). A second strand recommends against reliance on hypothetical responses (Diamond and Hausman 1994; Hausman 2012), primarily on the grounds that they suffer from systematic biases and are often inaccurate (List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005). A third strand suggests that, because hypothetical biases are systematic, the responses may be good *predictors* of real choices even if they are not good predictions. Some studies raise the possibility of correcting such biases by estimating relationships involving both real and hypothetical choices (e.g., Blackburn, Harrison, and Rutström 1994, Fox et al. 1998, List and Shogren 1998, List and Shogren 2002, and Mansfield 1998), but associated methods for recovering treatment effects encounter an important tension.[1] On the one hand, if a policy or intervention has not been implemented or is rare, the analyst cannot account for factors that might make hypothetical bias larger or smaller for that treatment state. On the other hand, if real data are rich enough to learn the relationship between real and hypothetical choices for each treatment state, one might be able to measure treatment effects without hypothetical data.

This paper builds on the third strand. We develop methods for estimating treatment effects that can account for the factors that create differences in hypothetical bias across treatment states even when treatment variation in real choice data is absent, limited, or endogenous. Our methods work by exploiting variation in hypothetical evaluations and real choices across a set of similar choice settings, in combination with the differences between hypothetical evaluations for actual versus counterfactual treatments within each setting. After providing formal econometric theory, we test the method and its assumptions in a field application as well as simulated applications based on laboratory data.

A simple hypothetical example illustrates both the issues that motivate our investigation and the essence of our approach. Suppose a charitable organization that sponsors an annual fundraising event engages an analyst to evaluate a potential strategy for boosting

[1] To be clear, much of the stated preference literature focuses on estimating willingness-to-pay for non-market goods rather than on measuring treatment effects on decisions people actually make (our focus).

attendance: publicizing a major donor's commitment to match all ticket revenue. The analyst might consider gathering real attendance data for many diverse fundraising events (i.e., similar choice settings), including ones that have deployed this treatment. Suppose for the moment that this real-choice strategy is foreclosed because the available dataset does not include any events where a donor matched ticket revenue. As an alternative, the analyst might conduct a survey of potential attendees, asking each hypothetically whether they would buy tickets in one of two scenarios, either with the match or without. The difference between the hypothetical choice frequencies in these two scenarios is a valid estimate of the treatment effect if hypothetical bias is absent or invariant across treatments.[2] Unfortunately, there is strong evidence that hypothetical bias varies across goods and contexts (see, e.g., Fox et al. 1998 and List and Shogren 1998). Consequently, there is reason to expect that it also varies across treatments, and our applications confirm that expectation. With existing methods, there is no empirical basis for drawing inferences about that variation in such applications.

To resolve this difficulty, our method would use data on multiple fundraising events in a different way. Assuming as before that we observe no events where a donor matched ticket revenue, variation across the events does not allow us to infer *directly* how hypothetical bias varies across the treatment states of interest. However, we may be able to make such inferences *indirectly*. Specifically, the bias presumably depends on various motivational factors. For each fundraising event and treatment state, imagine eliciting a collection of hypothetical responses $\boldsymbol{H}$ that plausibly span the important factors. For example, in addition to asking about the choice the respondent would make, one might also ask about the choices the respondent thinks others would make. The difference between those responses potentially sheds light on motivations related to image and ego. One can also ask questions designed to illuminate specific motivations, such as whether the respondent considers a given action socially appropriate or ethically commendable. Critically, because fundraising events differ in many ways, the elements of $\boldsymbol{H}$ likely vary across them for reasons unrelated to the treatment. Therefore, one can recover the relationship between hypothetical bias and $\boldsymbol{H}$ from variation across events. Assuming there is nothing special about the variation in motivations that would result from changes in the treatment, the same setting in a different treatment state is, effectively, just another setting. One can

[2] Under the assumption that hypothetical bias is unrelated to the treatment, the analyst could also compute an adjustment factor based on real and hypothetical choices for the prevailing scenario, in which there is no match, and apply it to hypothetical choices for the counterfactual scenario, in which there is a match. As discussed in Section 2.4, some existing stated-preference methods involving "calibration" and "data enrichment" (see, e.g., Louviere, Hensher, and Swait 2000) take this approach.

therefore use the estimated relationship to infer the effect of the treatment on hypothetical bias through its observed effect on $\boldsymbol{H}$.

Section 2 formalizes a general version of this method. The applications we consider belong to the following class: people make choices in a variety of related settings (such as fundraising events), and there are at least two treatment states, $\underline{w}$ and $\overline{w}$ (such as the presence or absence of a match). For each setting j, we observe a real-choice outcome (such as a purchase frequency), $Y_j(w)$, for a single treatment state, W_j. We begin by introducing an assumption called *Invariant Mapping*, which formalizes the preceding intuition. Conceptually, the assumption posits that, if hypothetical bias varies across treatment states, its statistical relationship to the full vector of hypothetical responses $\boldsymbol{H}(w)$ is nevertheless stable. We identify the types of applications for which this assumption may be reasonable as well as those for which it likely fails. For applications in which treatment variation is unconfounded, the Invariant Mapping assumption justifies a simple two-step estimator: first regress $Y_j(W_j)$ on $\boldsymbol{H}_j(W_j)$ using cross-setting variation, then use the fitted equation to impute real choices and difference across treatments.[3] We demonstrate that this estimator is consistent for the average treatment effect under specified conditions and characterize its asymptotic distribution. Significantly, this result covers applications for which reduced-form real-choice methods are inapplicable either because there is no variation in the observed treatment state (as in our opening example), or because none of the treatment states for which the analyst observes real choices resembles the treatment of interest.

Even when the analyst observes some real choices in the treatment state of interest, there are two important classes of applications for which our method offers potential advantages. First, the treatment of interest (or its absence) may be relatively rare, in which case real-choice estimates may be imprecise. In contrast, because the Invariant Mapping assumption allows us to estimate a single model of the outcome as a function of hypothetical evaluations using *all* settings (treated and untreated), our procedure can still yield precise estimates. Second, treatment may be endogenous and the analyst may be unable to identify good instruments or helpful discontinuities. Our approach may nevertheless work well because the first step recovers the relationship between Y and $\boldsymbol{H}$ using *all* variation in $\boldsymbol{H}$—not only

[3]When $\boldsymbol{H}_j(W_j)$ includes a hypothetical choices outcome, $Y_j^H(w)$, that is directly comparable to $Y_j(W_j)$, this procedure is equivalent to running a regression of $Y_j(W_j) - Y_j^H(W_j)$ on $\boldsymbol{H}_j(W_j)$, then using that regression to infer hypothetical bias for both treatment states, and finally computing treatment effects by calculating the difference between $Y_j^H(\overline{w})$ and $Y_j^H(\underline{w})$ after adjusting each for the implied hypothetical bias. The two-step formulation is more general because it encompasses applications in which the choice outcome of interest does not have a direct hypothetical counterpart, such as in our field application.

variation associated with treatment states, but also all other "ambient" variation (in our opening example, features of the fundraising events such as the charitable cause, venue, entertainment, etc.). Endogeneity contaminates the estimated relationship only through the former. We therefore obtain a limiting result: when the ambient variation is sufficiently important relative to the treatment variation, our two-step approach is approximately consistent. We also identify some simple diagnostics that can help the analyst gauge the relative importance of ambient variation. Further, we propose two estimation strategies that have desirable properties even if the limiting result for the two-step procedure does not apply. For the first strategy, we reformulate the endogeneity problem as a sample selection issue and derive an estimator that corrects for selection. An important feature of this method is that identification need not rely on distributional assumptions such as joint normality of the unobservables. For the second strategy, we derive bounds on the treatment effect without making any assumptions on treatment assignment. It turns out that the Invariant Mapping assumption can, by itself, substantially shrink the identified set regardless of how treatment is assigned.

An additional advantage of our approach is that it permits more flexible exploration of heterogeneous treatment effects than standard real-choice methods. Specifically, one can construct estimates for any subset of settings, even if there is no variation in the observed treatment state within that subset. In applications with endogenous treatment variation, one can calculate effects such as the ATE and the ATT.

We provide proof of concept by using the method to assess the effects of matching provisions on lending through a microfinance platform (Section 3). The observational data tell us the speed at which each borrower profile attracted funding and whether a third party offered matching funds. However, due to endogeneity of the match, observational data alone yield a biased estimate of the ground truth treatment effect, which we infer from a controlled experiment. Estimates of treatment effects that use our methods are reasonably close to ground truth.

Finally, in Section 4, we address a broader range of issues concerning the performance of our method by simulating applications using data from a laboratory experiment. The objective in each simulated application is to estimate the price responsiveness of demand for a collection of snack food items. Settings are snack food items, treatments are prices, and outcomes are purchase frequencies. We assume the analyst observes the demand for each item at a single price. The main advantage of the experiment is that we *actually* observe the demand for each item at more than one price. This feature enables us to test

whether the Invariant Mapping assumption holds in these data. It also allows us to simulate applications with treatment selection processes that our field application does not cover, including ones in which observed treatment variation is absent (meaning that the price is the same for all items), limited (similar to the field application examined in Whitehead, Weddell, and Groothuis 2016), or random. Additionally, this feature allows us to derive a ground truth estimate of price responsiveness for each item and use these estimates to evaluate the accuracy with which our approach captures the variation in treatment effects across settings. Results are once again generally favorable.

To be clear, hypothetical responses have limitations, and our approach is only applicable under the conditions we identify. As we explain, the assumptions that justify it are potentially problematic in other contexts, such as applications involving outcomes other than human choices. It may be difficult to depict complex and nuanced decision problems faithfully in a survey, or to survey populations that sufficiently resemble the decision makers. The approach also requires the analyst to obtain observational data on real choices, as well as survey data on hypothetical evaluations, for a variety of decision settings. Nonetheless, our analysis suggests that, in some applications, combining hypothetical and real data offers potentially significant advantages.

2 Method

2.1 The problem

Our objective is to estimate the effect of a treatment $w \in \mathbb{W}$ on choices made in various settings indexed by $j = 1, \ldots, J$. For each setting j, the potential outcome $Y_j(w)$ represents an aggregation of people's choices (e.g., a sum or average) given treatment w. The average treatment effect (ATE) comparing two treatment levels $\overline{w}$ and $\underline{w}$, both from $\mathbb{W}$, is

$$\tau_{\overline{w},\underline{w}} = \mathbb{E}\big(Y(\overline{w}) - Y(\underline{w})\big),$$

where the expectation is taken over a population of settings.

For concreteness, we preview the two applications in this paper:

Matching of charitable contributions. The analyst seeks to estimate the potentially heterogeneous effects of matching provisions for contributions to appeals posted on an online platform. Here, settings correspond to appeals, the treatment is the existence of a

match, and the outcomes are donation decisions by the platform's users.[4]

Product demand. The analyst seeks to estimate price elasticities for a collection of products (alternatively, for the same product across different markets). Here, settings correspond to products (alternatively, markets), the treatment is price, and outcomes are purchase decisions by customers.

We selected these two applications because they allow us to compare the results of our method against ground truth casual effect estimates from real choice experiments. Other examples of potential applications that fit our framework well include measuring how job applicants' choices over employers respond to workplace attributes (as in Wiswall and Zafar 2018), employers' interview decisions respond to features of job applicants' resumes, student applications and enrollment respond to school attributes, and employees' saving responds to pension plan features.

In this section, we consider various methods for estimating $\tau_{\overline{w},\underline{w}}$ using hypothetical evaluations with and without observations of real choices, and identify theoretical conditions under which the estimators have desirable properties and tractable distributions. Throughout, we use the following notation. The vector $\boldsymbol{H}_j(w)$ contains hypothetical evaluations of setting j under treatment w, aggregated over a sample of respondents for each (j, w). Ideally, the elicitations employ the same procedures and protocols for all j and w, and the samples are all representative of the same respondent population.[5] Ordinarily (but not necessarily), $\boldsymbol{H}_j(w)$ includes an aggregated measure of hypothetical choices $Y_j^H(w)$, which may or may not be directly comparable to the aggregate choice outcome $Y_j(w)$; our applications provide examples of both possibilities. $\boldsymbol{H}_j(w)$ may also incorporate other subjective evaluations such as reactions to the setting that incline people to choose particular types of options or to misstate the choices they would likely make. If real choices are observed, there is a single realized treatment state W_j for each setting j; i.e., $Y_j = Y_j(W_j)$. Settings may have fixed characteristics $\boldsymbol{X}_j$ pertaining to the nature of the decision problems or the populations who make the real choices. For simplicity, we assume that $(\boldsymbol{H}_j(W_j), \boldsymbol{H}_j(\overline{w}), \boldsymbol{H}_j(\underline{w}), Y_j, W_j, \boldsymbol{X}_j)_{j=1}^J$ is a random sample of i.i.d. settings j unless noted otherwise.

[4] For similar applications, see Karlan and List (2007) and Huck and Rasul (2011).

[5] Implicitly, the samples are large enough so that differences in the composition of respondents across (j, w) pairs average out. We relax this assumption below under the heading of "measurement error."

2.2 The relationship between real choices and hypothetical evaluations

To learn about real choices from hypothetical evaluations, one must assume the existence of a relationship between the two. We define the mapping from hypothetical evaluations to real outcomes in treatment state w as

$$\mu_w(\boldsymbol{h}, \boldsymbol{x}) = \mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}).$$

Our objective is to identify assumptions on this mapping that allow us to recover it and draw inferences about $\tau_{\overline{w},\underline{w}}$.[6]

2.2.1 No hypothetical bias

A possible strategy for estimating $\tau_{\overline{w},\underline{w}}$ is to ask people what they would choose in each setting and treat their responses as if they are real choices. Formally, suppose the outcome of interest is the average value of some choice variable. In that case, by calculating the average of the values respondents say they would choose, we can construct a measure of $Y^H(w)$ that is directly comparable to $Y(w)$. We can then estimate average treatment effects by computing the difference between average hypothetical choices in each treatment state:

$$\hat{\tau}^{simple}_{\text{hyp}} = \overline{Y^H(\overline{w})} - \overline{Y^H(\underline{w})},$$

where $\overline{Y^H(w)} = \frac{1}{J}\sum_{j=1}^{J} Y_j^H(w)$ is the sample average of the hypothetical choice under treatment state w for all settings.[7] This strategy is justified under the strong assumption that, in response to an appropriately worded hypothetical decision task, people report the choices they would actually make without systematic error: $Y_j^H(w) = Y_j(w) + \epsilon_j(w)$ with $\mathbb{E}(\epsilon(w) \mid Y^H(w)) = 0$. In that case, the mapping from hypotheticals to real choices simplifies as follows:

$$\mu_w^{simple}(h) = h$$

[6] In some applications, $\boldsymbol{H}$ may include a hypothetical choice variable $Y^H(w)$ that is directly comparable to $Y(w)$. Instead of recovering μ_w, one could then attempt to recover the relationship governing expected hypothetical bias, $b_w(\boldsymbol{h}, \boldsymbol{x}) = \mathbb{E}(Y(w) - Y^H(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})$, and use the estimated relationship to adjust observed choices. That approach is equivalent to ours but adds a step, and is also less general because it requires the existence of a directly comparable hypothetical choice.

[7] For example, Wiswall and Zafar (2018) estimate the effects of workplace amenities $w \in \mathbb{W}$ on hypothetical probabilities of choosing jobs $Y_j^H(w)$ in different scenarios j.

given the hypothetical choice $h = Y^H(w) = \boldsymbol{H}(w)$, regardless of fixed characteristics $\boldsymbol{X}$.

If this method worked, it would have several desirable properties. First, one could investigate the effects of treatments that have not yet been implemented. Second, because it is easy to elicit both treated and untreated hypothetical choices for all settings, no difficulties would arise even if in the real world treatments were assigned endogenously. Third, in addition to estimating the ATE for the entire sample, one could also estimate heterogeneous treatment effects by calculating $Y_j^H(\overline{w}) - Y_j^H(\underline{w})$ for arbitrary subsamples of settings.

Unfortunately, existing evidence overwhelmingly suggests that hypothetical choices are biased measures of real choices (List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005). With a constant bias ($\mu_w(Y^H(w)) = Y^H(w) + b$ for some fixed parameter b), $\hat{\tau}_{\text{hyp}}^{simple}$ would still be consistent.[8] However, as our applications illustrate, the simple estimator performs poorly, and a fixed bias (in levels or logs) is often implausible.

Researchers have attempted to reduce bias by changing how hypothetical choices are framed (for example, Cummings and Taylor 1999; Jacquemet et al. 2013), but it is unclear whether any such approach can provide a reliable general solution. Notably, in one of our applications, we find that these alternatives sometimes appear to reduce bias only because they introduce additional biases that are fortuitously counterveiling.

2.2.2 Proceeding without further assumptions

At the other end of the spectrum, we could attempt to learn about real treatment effects from hypothetical choices without imposing any assumptions on $\mu_w(\boldsymbol{h}, \boldsymbol{x})$. If this mapping were known, one could determine the ATE by comparing average fitted values across treatment states:

$$\hat{\tau}_{\text{hyp}}^{unrestricted} = \overline{\mu_{\overline{w}}} - \overline{\mu_{\underline{w}}},$$

where $\overline{\mu_w} = \frac{1}{J}\sum_{j=1}^{J} \mu_w(\boldsymbol{H}_j(w), \boldsymbol{X}_j)$ is the sample average of the "fitted" real choice under treatment state w for all settings.

This generalization of the simple approach has two important features. First, knowledge of the mapping μ_w allows the analyst to correct for hypothetical choice bias. Figure 1(a) illustrates this point for the case in which $\boldsymbol{h}$ consists of a single hypothetical choice variable $Y_j^H(w)$ that is directly comparable to $Y_j(w)$. If it were the case that $\mu_w(Y^H(w)) = Y^H(w) + b$

[8] To estimate a treatment effect that varies proportionately with the level of Y, and to allow for a proportional bias in hypothetical choices, one would simply reinterpret $Y(w)$ and $Y^H(w)$ as logs of the outcome of interest.

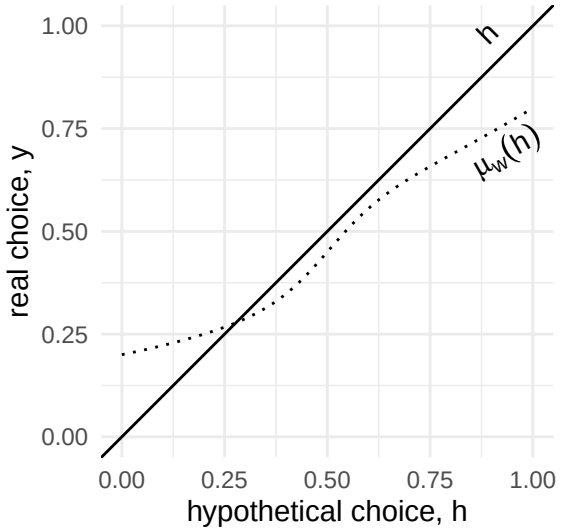
for some fixed parameter b, μ_w would be parallel to the 45 degree line. However, in many settings, the shape of μ_w is more complex, as shown by the dotted curve. If we could recover this mapping separately for each treatment state, we could undo the hypothetical choice bias.

Second, because the generalized mapping treats hypothetical evaluations as *predictors* rather than as *measures* of choices, its arguments can include multiple types of hypothetical evaluations. One implication of this observation is that the analyst does not need to construct a hypothetical choice aggregate, $Y^H(w)$, that is measured on the same scale and hence directly comparable to the choice outcome $Y(w)$; see our microfinance application for an illustration. Another implication is that the analyst may use any subjective response that aids prediction. $\boldsymbol{H}_j$ can thus include not only hypothetical choices Y_j^H (which are aggregates of multiple underlying motivations), but also measures of specific motivations, such as the extent to which a given option satisfies a desire for health, as well as measures that may predict the direction and magnitude of hypothetical choice bias, such as whether a given option is considered socially virtuous.

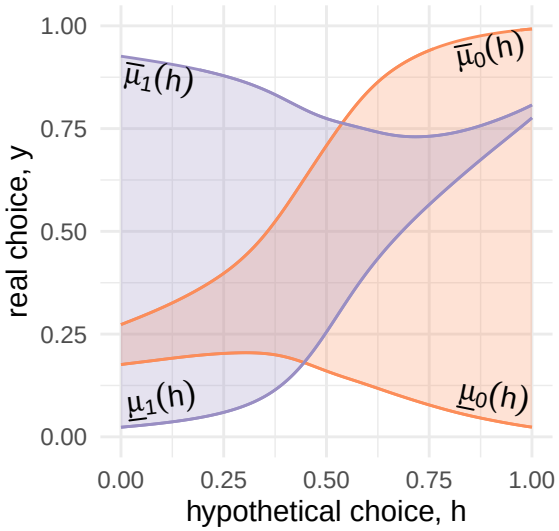
Unfortunately, even when the analyst observes both hypothetical and real choices, estimation of the unrestricted mapping can be challenging if real choices are observed in only a single treatment state $W_j \in \mathbb{W}$ for each setting j. Here we distinguish between three classes of applications according to whether treatment variation is absent, random, or endogenous. One cannot use the generalized approach when treatment variation is absent because μ_w is only recoverable for observed treatments. The approach is likewise problematic with endogenous treatments because differences in the observed relationship between real choices and hypothetical evaluations across treatment states could be attributable either to differences between $\mu_{\overline{w}}$ and $\mu_{\underline{w}}$ or to treatment selection. Following Manski (1990), one can derive bounds on the functions $\mu_{\overline{w}}$ and $\mu_{\underline{w}}$, but the resulting degree of indeterminacy for treatment effects may well render the analysis uninformative (as in Figure 1(b), where the bounds on each individual function are wide). Consequently, without further assumptions, applications are limited to the second class. But if one observed real choices under random treatment assignment, one could also estimate $\tau_{\overline{w},\underline{w}}$ without hypothetical evaluations.

2.2.3 The Invariant Mapping assumption

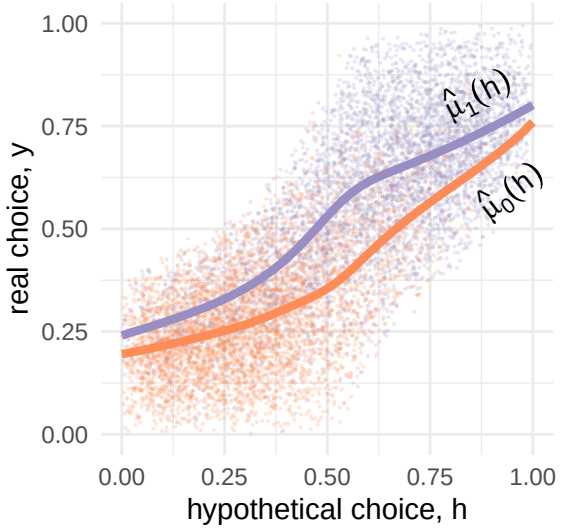
The alternatives considered so far involve assumptions that are either too strong to trust or too weak to support useful inferences. This paper formulates an intermediate alternative. Specifically, our central assumption is that the mapping relating hypothetical evaluations to



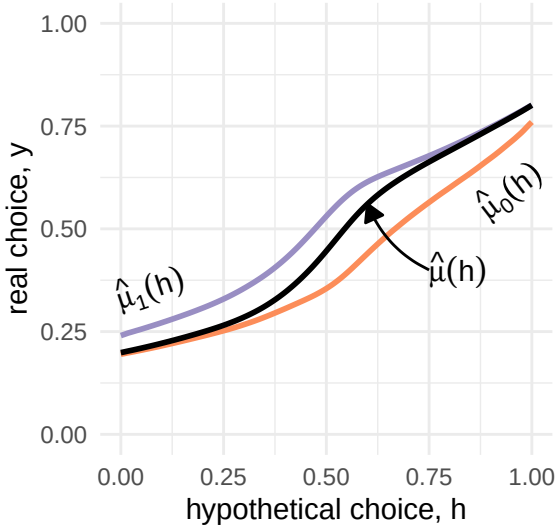
(a) True relationship between real and hypothetical choices ($y = \mu_w(h)$) compared with standard assumption ($y = h$)



(b) Estimated Manski (1990) bounds on true relationship



(c) Relationships between $Y(w)$ and $H(w)$ in observable data (conditional on $W = w$) where $\mu_0(h) = \mu_1(h)$



(d) Our method obtains a point estimate of μ

Figure 1: Example of identification and estimation with invariant mapping and binary treatment.

real choices, $\mu_w(\boldsymbol{h}, \boldsymbol{x})$, is the same for all treatment states. Formally:

Assumption 1. Invariant Mapping. *There exists a function* $\mu : \mathbb{H} \times \mathbb{X} \to \mathbb{R}$ *such that for all* $w \in \mathbb{W}$, $\boldsymbol{h} \in \mathbb{H}$, *and* $\boldsymbol{x} \in \mathbb{X}$: $\mu_w(\boldsymbol{h}, \boldsymbol{x}) = \mu(\boldsymbol{h}, \boldsymbol{x})$.

Why might Assumption 1 hold? One possibility is that people are *mental statists*, meaning that they care only about the bundle of internal hedonic sensations each option yields. As long as preferences over these "sensation bundles" are stable, there will be a stable mapping that identifies the chosen alternative for any menu of anticipated sensation bundles. Invariance of the mapping $\mu(\boldsymbol{h}, \boldsymbol{x})$ would then follow provided the relationship between the underlying anticipated sensations and $\boldsymbol{h}$, the vector of hypothetical evaluations, is also stable. The latter property is plausible when (1) $\boldsymbol{h}$ includes a rich set of hypothetical evaluations that span the pertinent hedonic sensations,[9] and (2) respondents are reasonably self-aware and, with respect to the sensations they expect each option to yield, resemble those who make the real choices. Significantly, the invariance property can hold even if people systematically misreport their hypothetical evaluations (provided $\boldsymbol{h}$ also spans motives for misreporting), or if the evaluations reflect overlapping composites of anticipated sensations.[10] The use of such composites avoids the need to fully catalog motivational attributes and to pair each with a matching proxy. Under these assumptions, there is nothing special about the portion of variation in $\boldsymbol{H}$ that is attributable to w: the treatment is just one of many factors that create variation in anticipated hedonic mental states, which in turn drive choice.

One important implication of the assumption is that the hypothetical evaluations capture the causal effect of the treatment.[11] This property is plausible if respondents are able to anticipate key factors that determine the effects of the treatment, even if only subcon-

[9] One does not actually need $\boldsymbol{h}$ to subsume *all* sensations that affect observed choices. A natural possibility is that people answer hypothetical questions by envisioning the *typical* conditions under which such a decision would be made, such as how eager one typically is to support a particular charity, rather than the *specific* conditions that give rise to the observed value of Y, such as the appearance of a news story that makes one especially inclined to provide that support. As long as the idiosyncratic effects of these specific conditions are orthogonal to the information contained in $\boldsymbol{H}$ as well as to the treatment W, this consideration simply adds randomness to the relationship between Y and $\boldsymbol{H}$ without overturning Assumption 1. See Appendix B for an elaboration of this possibility.

[10] To illustrate, suppose $\boldsymbol{s}$ is an N-dimensional vector of the pertinent hedonic sensations and $\boldsymbol{h}$ is an M-dimensional vector of reported hypothetical evaluations, where $M \geq N$. Suppose further that anticipated sensations fully govern hypothetical responses: $\boldsymbol{h} = f(\boldsymbol{s})$. Let $\mathbb{S}$ denote the domain of potential sensations, and let $\mathbb{H} = f(\mathbb{S})$. Then the existence of a stable inverse mapping f^{-1} on the set $\mathbb{H}$ is a sufficient condition for being able to use $\boldsymbol{h}$ as a "stand in" for $\boldsymbol{s}$. Significantly, this formulation allows for systematic misreporting, including the possibility that certain sensations influence the accuracy with which others are reported.

[11] Notably, the assumptions underlying the simple approach have the same implication.

sciously.[12] While people may be quite good at anticipating the types of mental reactions that will influence their choices, they may do less well when anticipating the causal factors behind an outcome that depends on other processes. For example, if one were interested in measuring the impact of water purification policies on health, one could ask respondents to forecast health outcomes under various hypothetical policies. However, their limited understanding of the underlying biological processes would likely cause the relationship between health outcomes and forecasts to vary across policies. Similarly, if the sample of individuals who provide hypothetical evaluations differs too much from the population of decision makers, their responses may fail to capture the effects of the treatment on the anticipated hedonic sensations that drive observed choices.[13] Similar problems may arise if the hypothetical scenarios are either insufficiently familiar or too complex to fully represent when gathering responses. For example, hypothetical automobile purchase decisions typically cannot include test drives.

Assumption 1 may also fail if $\boldsymbol{H}(w)$ is not rich enough to adequately span the set of hedonic sensations governing both choice and misreporting.

2.3 Estimation

Next we turn to estimation under the Invariant Mapping assumption. We begin by considering the first two classes of applications, which share the property that the treatment is unconfounded, in the first case because it has no variation and in the second because it has random variation. Then we consider the third class, focusing on applications in which treatment is confounded conditional on the hypothetical responses and fixed characteristics.

For many results in this section, we assume for simplicity that the conditional expectation function is linear. We state this assumption in combination with mapping invariance:

Assumption 2. Linearity. *The conditional expectations of potential outcomes are linear in the*

[12] Significantly, hypothetical choices exhibit many of the same anomalies as real choices; see Poe (2016).

[13] This consideration argues for sampling respondents from the population that makes choices, with minimal temporal separation. However, doing so raises two issues. First, in applications where the population of decision-makers varies across settings, the population of respondents would also vary. In that case, the method implicitly assumes that these populations differ across settings only with respect to their mappings from options to anticipated sensations: μ may not be stable unless the mappings from anticipated sensations to choices and reports are similar across populations. A related assumption appears in Bernheim, Kim, and Taubinsky (2024). Second, respondents' real choices may distort their hypothetical evaluations (for example, through anchoring or ex post rationalization). This possibility is less of a concern for decisions that are differentiated or forgettable; see our microlending application for an example. In some applications, it may be possible to mitigate the concern by eliciting hypothetical evaluations prior to the treatment's implementation, or by identifying a similar but unexposed subpopulation.

predictors: for $w \in \mathbb{W}$*,*

$$\mathbb{E}\Big(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}\Big) = \boldsymbol{h}\boldsymbol{\beta} + \boldsymbol{x}\boldsymbol{\gamma} \tag{1}$$

Relaxing this assumption is straightforward. Even with this specification, one can account for nonlinearities by including higher order terms and interactions in the vectors $\boldsymbol{h}$ and $\boldsymbol{x}$.

Linearity is a particularly appealing approximation if $\boldsymbol{h}$ includes a hypothetical choice variable Y^H that is directly comparable to the choice outcome Y. In that case, we can rewrite equation (1) as $\mathbb{E}\Big(Y(w) - Y^H(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}\Big) = \boldsymbol{h}\boldsymbol{\delta} + \boldsymbol{x}\boldsymbol{\gamma}$,where all elements of δ and β coincide except for the coefficient of Y^H. Expressing the relationship this way highlights that our formulation allows the hypothetical bias, $Y_j(w) - Y_j^H(w)$, to vary with hypothetical responses capturing features of the choice task that potentially motivate misreporting.

2.3.1 Estimation with unconfounded treatments

When treatment assignment is random, the observed relationship between real choices and hypothetical evaluations does not depend on treatment status. Consequently, to obtain consistent estimates of the mapping's parameters ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$), one can regress $Y_j(W_j)$ on $\boldsymbol{H}_j(W_j)$ and $\boldsymbol{X}_j$ using the observations in any given treatment state, $W_j = w$. Alternatively, to maximize power, one can run the same regression pooling data across all treatment states because the parameters are the same under the Invariant Mapping assumption.

These observations motivate the following two-step procedure for estimating the ATE (Step 1 being the regression suggested at the end of the previous paragraph):[14]

Step 1. Using data for the realized treatment states, estimate the relationship between choices and the corresponding hypothetical evaluations:

$$Y_j = \boldsymbol{H}_j\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \epsilon_j, \tag{2}$$

where $Y_j = Y_j(W_j)$ is the realized outcome, hypothetical evaluations $\boldsymbol{H}_j = \boldsymbol{H}_j(W_j)$ correspond to the realized treatment state W_j, and $\boldsymbol{X}_j$ is a collection of observable, fixed characteristics (including the intercept).

[14]For binary treatments ($\overline{w} = 1$, $\underline{w} = 0$), one can equivalently recover the average treatment effect by estimating the regressions $Y_j(W_j) = \boldsymbol{H}_j(W_j)\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \epsilon_j(W_j)$ and $\boldsymbol{H}_j(w)^T = w\boldsymbol{\delta} + \boldsymbol{\nu}_j(w)^T$, and then computing $\tau = \boldsymbol{\delta}'\boldsymbol{\beta}$.

Step 2. Use the estimated relationship to predict outcomes for both states, and take the difference:

$$\hat{\tau}_{\overline{w},\underline{w}} = \big(\overline{\boldsymbol{H}(\overline{w})} - \overline{\boldsymbol{H}(\underline{w})}\big)\hat{\boldsymbol{\beta}}$$

where $\overline{\boldsymbol{H}(w)} = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{H}_j(w)$ is the sample average of the predictors under treatment state $w \in \{\underline{w}, \overline{w}\}$ for all settings.[15]

Notice that the procedure's logic only requires treatment assignment to be random *conditional* on $\boldsymbol{H}_j$ and $\boldsymbol{X}_j$. Consequently, the estimator should also perform well in settings with endogenous treatments under the strong assumption that hypothetical responses (along with other control variables) capture all the choice-relevant factors that influence treatment selection.[16] The following assumption covers these possibilities along with random assignment:

Assumption (3A)**. Unconfoundedness.** *Treatment assignment is unconfounded conditional on hypothetical evaluations*: for $w \in \mathbb{W}$, $W \perp\!\!\!\perp Y(w) \mid \boldsymbol{H}(w), \boldsymbol{X}$.

Notice also that the procedure's logic applies equally well in settings where there is no treatment variation. When all observed choices correspond to a single treatment state, say w, we can estimate μ_w consistently. Then, under the Invariant Mapping assumption, we can use the estimated relationship to project choices in other treatment states. Technically, this possibility is a special case of unconfoundedness but, given its importance, we state it separately:

Assumption (3B)**. No treatment variation.** *There is no variation in realized treatment:* $W_j = w$ *a.s. for all settings* j *and some* $w \in \mathbb{W}$.

Either of these assumptions is sufficient to ensure that our linear estimator is consistent and has a tractable asymptotic distribution.

Proposition 1. *Suppose the data* $(\boldsymbol{H}_j(W_j), \boldsymbol{H}_j(w), \boldsymbol{H}_j(\underline{w}), Y_j, W_j, \boldsymbol{X}_j)_{j=1}^{J}$ *are a random sample of independent observations and standard regularity conditions hold. Under Assumptions 1*

[15] The use of hypothetical data has been criticized on the grounds that it does not necessarily satisfy an "adding-up test;" see, e.g., Diamond and Hausman (1994). Significantly, our method always satisfies this test. Specifically, let w_0 represent the status quo and w_i represent treatment state $i \in \{A, B, AB\}$, where A and B are interventions and AB indicates that both are in effect. Then, mechanically, $\hat{\tau}_{w_0,w_A} + \hat{\tau}_{w_A,w_{AB}} = \hat{\tau}_{w_0,w_B} + \hat{\tau}_{w_B,w_{AB}} = \hat{\tau}_{w_0,w_{AB}}$.

[16] In Appendix B, we provide an explicit model of a plausible underlying process for which Assumption 3A holds even though treatment is endogenous and hypothetical evaluations do not capture all the conditions contributing to the choice aggregate in each setting.

and 2, if either Assumption 3A or 3B holds, the parametric estimator $\hat{\tau}_{\overline{w},\underline{w}}$ is consistent for the average treatment effect $\tau_{\overline{w},\underline{w}}$ and asymptotically normal:[17]

$$\sqrt{J}\Big(\hat{\tau}_{\overline{w},\underline{w}} - \tau_{\overline{w},\underline{w}}\Big) \to \mathcal{N}\Big(0, V_\tau\Big).$$

Overlap. The approach uses observed variation across settings (and possibly treatments) to estimate the relationship between $\boldsymbol{H}$ and Y, and then predicts the treatment's effect on Y based on how it shifts $\boldsymbol{H}$. For applications in which the counterfactual treatment states shift the distribution of $\boldsymbol{H}$ substantially outside the range of observed variation, the latter step extrapolates based on a functional form assumption. Caution is warranted in such cases because the treatment effect is unidentified without that assumption. In contrast, when the distributions of $\boldsymbol{H}$ under the observed and counterfactual treatment states overlap, treatment effects are identified semiparametrically, and one can dispense with the functional form assumption entirely (as shown in Appendix C.2).[18] Overlap is easy to assess empirically by checking whether the support of the distribution of $(\boldsymbol{H}_j(W_j), \boldsymbol{X}_j)$ spans the support of $(\boldsymbol{H}_j(w), \boldsymbol{X}_j)$ for $w \in \{\underline{w}, \overline{w}\}$. When the treatment is binary, this condition is always satisfied when the support of $(\boldsymbol{H}_j(W_j), \boldsymbol{X}_j)$ spans that of $(\boldsymbol{H}_j(1-W_j), \boldsymbol{X}_j)$. It tends to hold when the effect of the treatment on evaluations is not too large relative to other sources of variation in evaluations across settings—a property that has other important implications

[17] The formula for the variance matrix is:

$$\begin{aligned} V_\tau = {} & \mathbb{E}\Big((\tau_{\overline{w},\underline{w}} - (\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w}))\boldsymbol{\delta})^2\Big) \\ & + \mathbb{E}\Big(\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w})\Big)\boldsymbol{V}^{\text{ols}}\mathbb{E}\Big(\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w})\Big)^T \\ & - 2\mathbb{E}\Big(\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w})\Big)\mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}^T(Y - \boldsymbol{Z}\boldsymbol{\delta})(\tau_{\overline{w},\underline{w}} - (\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w}))\boldsymbol{\delta})\Big), \end{aligned}$$

where, for notational convenience, we denote the full sets of regressors by $\boldsymbol{Z}_j(w) = [\boldsymbol{H}_j(w),\ \boldsymbol{X}_j]$, $\boldsymbol{Z}_j = \boldsymbol{Z}_j(W_j)$, and the joint vector of their coefficients by $\boldsymbol{\delta} = [\boldsymbol{\beta}^T,\ \boldsymbol{\gamma}^T]^T$. $\boldsymbol{V}^{\text{ols}} = \mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}(y - \boldsymbol{Z}\boldsymbol{\delta})^2\Big)\mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}\Big)^{-1}$ is the asymptotic variance matrix of the OLS estimator $\hat{\boldsymbol{\delta}} = [\hat{\boldsymbol{\beta}}^T,\ \hat{\boldsymbol{\gamma}}^T]^T$ from Step 1. The proof follows from writing the two-step estimator in the GMM framework (cf. Newey and McFadden 1994); see Appendix C.1 for details.

[18] Our notion of overlap is weaker than the standard overlap assumption, which requires that the probability of receiving a treatment conditional on observables (the propensity score) is always bounded away from 0 and 1. Both assumptions ensure that (asymptotically) there are observed real choices for every value of the predictors when forming predictions of potential outcomes. In standard settings, every value of the predictors must occur separately among settings with $W_j = \overline{w}$ and for settings with $W_j = \underline{w}$. In contrast, our overlap assumption (see Appendix C.2) requires that every value of the predictors occurs for *some* observations in their realized treatment state W_j, but it does not matter whether that treatment state is $\overline{w}$ or $\underline{w}$. For instance, it can hold even when there is no variation in the treatment, where the standard overlap assumption would fail.

discussed in the next section. When the condition is satisfied, our method works better in practice (as demonstrated in Section 4.2.1).

Measurement error. When survey samples are small and not easily expanded, sampling error in $\boldsymbol{H}_j(0)$ and $\boldsymbol{H}_j(1)$, which average over respondents, can potentially bias $\hat{\tau}_{\overline{w},\underline{w}}$. In Appendix C.3 we characterize scenarios under which this bias is negligible. For applications in which the bias may be significant, we also describe an estimator that addresses the problem by splitting the sample of respondents and using one subsample to construct instruments for the other.

Advantages. Our approach is potentially advantageous not only in applications with no observed treatment variation (as mentioned above), but also in those where variation in the characteristics of observed treatments does not span the treatments of interest, for instance because the potential treatment space is high-dimensional. In such cases, our method may facilitate low-cost exploration of potential interventions.[19]

Even in applications where unconfounded treatment variation *does* encompass the interventions of interest, the method offers two potential advantages over standard real-choice methods. First, while imbalance in the sizes of treatment groups can impair the precision of standard real-choice methods, it has no direct impact on the precision of our estimator. This feature is a consequence of the Invariant Mapping assumption, which allows us to estimate a single model of the outcome as a function of hypothetical evaluations using *all* settings, treated and untreated. Second, unlike standard real-choice methods, our approach enables one to estimate a treatment effect conditional on *any* observed fixed characteristic even if there is no treatment variation for settings with that characteristic.

2.3.2 Estimation with confounded treatments

When real choices are observed only under endogenous treatment states, recovering the mapping μ becomes challenging. Figure 1(c) illustrates a case in which the settings where people would choose higher values of Y tend to be treated. Because choice confounds treatment status, inferring μ from treated observations yields an overestimate (purple line),

[19]Consider, for example, the problem of designing a form for opting in or opting out of organ donations (Kessler and Roth 2012). Our approach allows one to evaluate at low cost how treatment effects vary with nuanced features of the intervention such as the wording, font size, color, and placement of the question. Specifically, once one obtains estimates of $\mu(\boldsymbol{h}, \boldsymbol{x})$, evaluating additional treatment variants is simply a matter of gathering the requisite hypothetical responses. In contrast, similar exploration of the treatment space using real-choice methods would ordinarily require the implementation of numerous costly field experiments.

and inferring it from untreated observations yields an underestimate (orange line). Even though $\mu_0(h) = \mu_1(h) = \mu(h)$, estimating the curves based on endogenously selected data would therefore create the false appearance that the Invariant Mapping assumption fails, and lead to biased estimates of choice.

In this section, we begin by asking how our two-step estimator $\hat{\tau}_{\overline{w},\underline{w}}$ performs when endogeneity confounds the treatment, and we identify a condition under which the resulting estimate is approximately consistent. Then we ask how one might proceed in settings where the approximation may not hold. Our methods permit the estimation of causal effects when it is difficult or impossible to identify a source of exogenous variation in the treatment such as an instrumental variable or discontinuity. They also potentially preserve the advantages involving precision and measurement of heterogeneous treatment effects mentioned at the end of the previous section.

Approximate consistency of two step estimator. Suppose we used the simple two-step estimator $\hat{\tau}_{\overline{w},\underline{w}}$ from the previous section in a setting where the treatment is confounded. In that case, we would be estimating the relationship $\hat{\mu}(\boldsymbol{h}, \boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{H}(W) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})$, which Figure 1(d) illustrates for the one dimensional case.

As a general matter, there is no reason to think $\hat{\mu}(\boldsymbol{h}, \boldsymbol{x})$ coincides with $\mu(\boldsymbol{h}, \boldsymbol{x})$. However, it may be close. In Step 1 of the two-step method, we identify the parameter vector $\boldsymbol{\beta}$ by relating variation in the outcome variable to variation in the hypothetical responses. Bias arises because part of the variation in hypothetical responses reflects endogenous treatment variation. But there is also "ambient" variation in the hypothetical evaluations, which helps identify the parametric relationship. The greater that ambient variation, the more it dilutes the endogenous variation associated with the treatment state. When the ambient variation is sufficiently important relative to the treatment variation, the bias becomes arbitrarily small. Thus, our method can identify causal effects accurately so long as most variation in hypotheticals is not from treatment.

To formalize this intuition, we maintain Assumptions 1 (Invariant Mapping) and 2 (Linearity), and focus on models that employ a single hypothetical evaluation without controlling for fixed characteristics. We then consider sequences of environments, indexed by k, such that as k grows the importance of the treatment shrinks *relative* to the ambient variation in hypotheticals:

Assumption (3C)**. Most variation not due to treatment.** *For at least one treatment state* $w \in \mathbb{W}$*: as* $k \to \infty$*, (i)* $\frac{\text{var}_k(H_j(W_j) - H_j(w))}{\text{var}_k(H_j(w))} \to 0$*, (ii)* $\frac{\text{var}_k(\epsilon_j(W_j) - \epsilon_j(w))}{\text{var}_k(H_j(w))} \to 0$*, and (iii)*

$\sqrt{\frac{\text{var}_k(\epsilon_j(w))}{\text{var}_k(H_j(w))}} < \infty$.

The assumption states that the ambient variation in hypotheticals is large relative to (i) the variation in hypotheticals that is attributable to the treatment, and (ii) the variation in the error term that is attributable to the treatment. Condition (iii) requires that the model linking outcomes to hypotheticals does not become arbitrarily poor in terms of fit as one moves to the limit (e.g., the R^2 in a regression of $Y_j(0)$ on $H_j(0)$ remains bounded away from 0).

Introducing this assumption yields the following result:

Proposition 2. *Suppose Assumptions 1, 2, and 3C hold,* $0 < \text{var}_k(H_j) < \infty$ *and* $|\text{cov}_k(H_j, Y_j)| < \infty$ *for all fixed* k*, and* $\mathbb{E}_k(H_j(1) - H_j(0))$ *is bounded. Then the asymptotic bias of* $\hat{\tau}_{\overline{w},\underline{w}}$ *vanishes, i.e.,* $\lim_{k\to\infty} \text{plim}_{n\to\infty} \hat{\tau}_{\overline{w},\underline{w},k} - \tau_{\overline{w},\underline{w},k} = 0$.[20]

As a special case, the key conditions (i) and (ii) are satisfied if treatment is rare (probability of treatment $p_k := \mathbb{E}_k(W_j) \to 0$ as $k \to \infty$) or very common ($p_k \to 1$ as $k \to \infty$), provided potential outcomes and hypotheticals are bounded and have non-zero variance.

The proposition suggests that one can potentially reduce endogeneity bias by collecting choice data from more diverse settings. For example, in our microfinance application, one could create more ambient variation by oversampling borrower profiles that are unusually desirable or undesirable according to the various subjective measures. However, there is a qualification: the model relating outcomes to hypotheticals must continue to perform well across the diverse settings. Combining very different types of choices, such as decisions about lending to microenterprises and large corporations, would make the settings more diverse but would likely undermine the model's predictive performance, which could violate condition (iii).

Diagnostics for Assumption 3C. In any given application, simple diagnostics based on conditions (i) and (iii) can help the analyst assess the relative amount of ambient variation in hypotheticals. Calculating the ratio in condition (i) is straightforward and we check it in our applications. For condition (iii), the R^2 for a regression of Y_j on H_j using observations in a single treatment state is reasonably suggestive.

[20] When $\tau_{\overline{w},\underline{w},k} \neq 0$ for all k, the proportional bias also vanishes even if $\tau_{\overline{w},\underline{w},k} \to 0$, i.e. $\lim_{k\to\infty} \text{plim}_{n\to\infty} \frac{\hat{\tau}_{\overline{w},\underline{w},k} - \tau_{\overline{w},\underline{w},k}}{\tau_{\overline{w},\underline{w},k}} = 0$. Accordingly, one can think of this proposition as covering cases where conditions (i) and (ii) of Assumption 3C hold either because the variation in the hypothetical responses grows or the treatment effect shrinks. See the proof in Appendix C.4.

Next, we consider what the analyst might do if there is reason to suspect that ambient variation in hypotheticals is insufficient to invoke Assumption 3C.

Addressing endogeneity through sample selection corrections. One possibility is to introduce additional assumptions that justify alternative estimation procedures. To motivate our specific proposal, imagine taking the subset of settings associated with a single treatment state and regressing outcomes on hypotheticals along with fixed characteristics. Treatment endogeneity potentially biases the coefficient estimates for such regressions, and hence for our method, only because it makes the selection into that sample non-random. Accordingly, in our framework, one can potentially address treatment endogeneity through sample selection corrections.

In Appendix C.5, we model endogenous treatment assignment using an approach similar to that of Heckman (1976; 1979). For a binary treatment, our proposed estimation procedure involves preserving Step 2 of the two-step method and replacing Step 1 with the following:

Step 1a. Using data on all settings, estimate a Probit regression:

$$\Pr(W = 1 \mid \boldsymbol{H}(0), \boldsymbol{H}(1), \boldsymbol{X}) = \Phi(\boldsymbol{H}(0)\boldsymbol{\alpha}_0 + \boldsymbol{H}(1)\boldsymbol{\alpha}_1 + \boldsymbol{X}\boldsymbol{\alpha}_x)$$

where Φ is the standard normal cdf, and form fitted values for the index $V_j = \boldsymbol{H}_j(0)\hat{\boldsymbol{\alpha}}_0 + \boldsymbol{H}_j(1)\hat{\boldsymbol{\alpha}}_1 + \boldsymbol{X}_j\hat{\boldsymbol{\alpha}}_x$.

Step 1b. Using data for the realized treatment states, estimate the relationship between choices and the corresponding hypothetical evaluations, including the two scalar covariates $\frac{(1-W_j)\phi(V_j)}{1-\Phi(V_j)}$ and $\frac{W_j\phi(V_j)}{\Phi(V_j)}$ where ϕ and Φ are the standard normal pdf and cdf:

$$Y_j = \boldsymbol{H}_j\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \frac{(1-W_j)\phi(V_j)}{1-\Phi(V_j)}\sigma_{\epsilon,0} + \frac{W_j\phi(V_j)}{\Phi(V_j)}\sigma_{\epsilon,1} + \nu_j,$$

where $Y_j = Y_j(W_j)$ is the realized outcome, hypothetical evaluations $\boldsymbol{H}_j = \boldsymbol{H}_j(W_j)$ correspond to the realized treatment state W_j, $\boldsymbol{X}_j$ is a collection of observable fixed characteristics (including the intercept), and $\frac{(1-W_j)\phi(V_j)}{1-\Phi(V_j)}$ and $\frac{W_j\phi(V_j)}{\Phi(V_j)}$ are scalar covariates correcting for selection.

We call the resulting estimator $\hat{\tau}^{\text{hypit}}_{\overline{w},\underline{w}}$. Steps 1a and 1b are analogous to the "Heckit" estimator for problems with selection, except that Step 1b pools information across treatment states based on the Invariant Mapping assumption. An attractive feature of this approach is that

both $\boldsymbol{H}(0)$ and $\boldsymbol{H}(1)$ plausibly enter into the selection equation in Step 1a, but only the term corresponding to the observed treatment appears in the outcome model for a given observation in Step 1b (through $\boldsymbol{H}$). Consequently, $\boldsymbol{H}_j(1-W_j)$ effectively serves as an instrument, so that identification does not rely exclusively on functional form.[21]

Bounding. If the assumptions required for the sample selection estimator $\hat{\tau}^{\text{hypit}}_{\overline{w},\underline{w}}$ are considered excessive, another possibility is to derive bounds on the treatment effect without introducing additional assumptions. It turns out that the Invariant Mapping assumption can, by itself, substantially shrink the identified set for $\tau_{\overline{w},\underline{w}}$, regardless of how treatment is assigned. Figure 1(b) shows separate bounds for $\mu_0(h)$ (orange region) and $\mu_1(h)$ (blue region) based on Manski (1990). Imposing Assumption 1 shrinks the identified sets for $\mu_0(h)$ and $\mu_1(h)$ to the *intersection* of these bounds (the region where orange and blue overlap). Additionally, if the effect of the treatment on hypothetical evaluations (and hence on choices) is small, $\tau_{\overline{w},\underline{w}}$ may be determined quite precisely regardless of the degree of indeterminancy for the mapping μ. In the extreme case where the treatment has no effect on hypothetical evaluations, the identified set for $\tau_{\overline{w},\underline{w}}$ shrinks to a point regardless of possible endogeneity. The method is therefore generally useful for testing the null hypothesis of no effect.

The following proposition formalizes and generalizes these ideas.

Proposition 3. *Suppose Assumption 1 holds and potential outcomes are bounded,* $y_{lb} \leq Y_j(w) \leq y_{ub}$ *for* $w \in \mathbb{W}$. *Then the sharp identified set for the average treatment effect is* $\tau_{lb} \leq \tau_{\overline{w},\underline{w}} \leq \tau_{ub}$ *where the bounds* τ_{lb} *and* τ_{ub} *depend on functions* $(p_w, \mu_{Y|w}, f_{\boldsymbol{H}(w),\boldsymbol{X}})_{w\in\mathbb{W}}$ *that are point identified at all relevant values.*[22] *The length of the identified set for* $\tau_{\overline{w},\underline{w}}$ *is* $\int_{\mathbb{X}}\int_{\mathbb{H}}(\min_{w\in\mathbb{W}}\{\mu_{w,ub}(\boldsymbol{h},\boldsymbol{x})\} - \max_{w\in\mathbb{W}}\{\mu_{w,lb}(\boldsymbol{h},\boldsymbol{x})\})|f_{\boldsymbol{H}(\overline{w}),\boldsymbol{X}}(\boldsymbol{h},\boldsymbol{x}) - f_{\boldsymbol{H}(\underline{w}),\boldsymbol{X}}(\boldsymbol{h},\boldsymbol{x})|\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$.

Appendix C.6.2 explains how to estimate these bounds under the Invariant Mapping and Linearity assumptions.

[21] The "inverse Mills ratio" form of the correction terms follows from a normality assumption, but similar semiparametric approaches (for instance, Newey 2009) are feasible for sufficiently large samples. For extensions to more than two treatment values, see, for instance, Lee (1983).

[22] Define $p_w(\boldsymbol{h},\boldsymbol{x}) = \Pr(W = w \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})$, $\mu_{Y|w}(\boldsymbol{h},\boldsymbol{x}) = \mathbb{E}(Y \mid W = w, \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = x)$ and the density of $(\boldsymbol{H}(w), \boldsymbol{X})$ as $f_{\boldsymbol{H}(w),\boldsymbol{X}}$. Let $\mu_{w,\text{lb}}(\boldsymbol{h},\boldsymbol{x}) = p_w(\boldsymbol{h},\boldsymbol{x})\mu_{Y|w}(\boldsymbol{h},\boldsymbol{x}) + (1 - p_w(\boldsymbol{h},\boldsymbol{x}))y_{\text{lb}}$ and$\mu_{w,\text{ub}}(\boldsymbol{h},\boldsymbol{x}) = p_w(\boldsymbol{h},\boldsymbol{x})\mu_{Y|w}(\boldsymbol{h},\boldsymbol{x}) + (1 - p_w(\boldsymbol{h},\boldsymbol{x}))y_{\text{ub}}$. Let $\mu_{\text{lb}}(\boldsymbol{h},\boldsymbol{x}) = \begin{cases}\max_{w\in\mathbb{W}}\{\mu_{w,\text{lb}}(\boldsymbol{h},\boldsymbol{x})\} & \text{if } f_{\boldsymbol{H}(\overline{w}),\boldsymbol{X}}(\boldsymbol{h},\boldsymbol{x}) > f_{\boldsymbol{H}(\underline{w}),\boldsymbol{X}}(\boldsymbol{h},\boldsymbol{x}) \\ \min_{w\in\mathbb{W}}\{\mu_{w,\text{ub}}(\boldsymbol{h},\boldsymbol{x})\} & \text{otherwise}\end{cases}$ and $\mu_{\text{ub}}(\boldsymbol{h},\boldsymbol{x}) = \begin{cases}\min_{w\in\mathbb{W}}\{\mu_{w,\text{ub}}(\boldsymbol{h},\boldsymbol{x})\} & \text{if } f_{\boldsymbol{H}(\overline{w}),\boldsymbol{X}}(\boldsymbol{h},\boldsymbol{x}) > f_{\boldsymbol{H}(\underline{w}),\boldsymbol{X}}(\boldsymbol{h},\boldsymbol{x}) \\ \max_{w\in\mathbb{W}}\{\mu_{w,\text{lb}}(\boldsymbol{h},\boldsymbol{x})\} & \text{otherwise}\end{cases}$. Then $\tau_{\text{lb}} = \int_{\mathbb{X}}\int_{\mathbb{H}}\mu_{\text{lb}}(\boldsymbol{h},\boldsymbol{x})(f_{(\boldsymbol{H}(\overline{w}),\boldsymbol{X})}(\boldsymbol{h},\boldsymbol{x}) - f_{(\boldsymbol{H}(\underline{w}),\boldsymbol{X})}(\boldsymbol{h},\boldsymbol{x}))\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$ and $\tau_{\text{ub}} = \int_{\mathbb{X}}\int_{\mathbb{H}}\mu_{\text{ub}}(\boldsymbol{h},\boldsymbol{x})(f_{(\boldsymbol{H}(\overline{w}),\boldsymbol{X})}(\boldsymbol{h},\boldsymbol{x}) - f_{(\boldsymbol{H}(\underline{w}),\boldsymbol{X})}(\boldsymbol{h},\boldsymbol{x}))\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$.

2.3.3 Extensions

In this section, we mention some potentially useful extensions of the approach.

Machine learning. When exploring robustness with respect to rich specifications that include many hypothetical variables along with transformations and interactions, the analyst may wish to guard against overfitting by deploying machine learning tools. Appendix C.7 describes an approach similar to LASSO building on approximate residual balancing (ARB, Athey, Imbens, and Wager 2018).

Respondent representativeness. In some applications, those answering hypothetical questions may differ from the people whose choices determine the real outcomes. For example, in our microfinance application, visitors to the website determine the outcome of interest, but we obtain hypothetical evaluations by drawing a sample of respondents from Amazon Mechanical Turk, fewer than 25% of whom report having visited the website. In Appendix C.8, we describe and implement an extension that uses leave-one-out measures of response predictiveness to identify the respondents most skilled at predicting real choices. Relying on those responses can in some cases improve performance.

2.4 Methodological precursors

Next we discuss the relationships between our method and other types of analyses that combine data on real choices with hypothetical responses.

Data enrichment. The existing literature on "data enrichment" also includes methods for recovering treatment effects. Typically, the analyst estimates a standard discrete-choice model at the individual level "stacking" observations of both real and hypothetical choices.[23] When there is limited treatment variation for real choices in the setting of interest, the hypothetical choices provide data for treatments that are either rare or unobserved. Significantly, such models can include adjustment terms that account for systematic differences between real and hypothetical observations, including hypothetical bias. When specifications involve multiple covariates, the analyst can test the appropriateness of simple adjustments. For

[23] For overviews, see Louviere, Hensher, and Swait (2000, ch. 8) and Whitehead et al. (2008). These methods are also referenced in the literature on "benefit transfer;" see, e.g., Johnston, Ramachandran, and Parsons (2015, ch. 9). For economic applications involving data enrichment, see Berry, Levinsohn, and Pakes (2004) and Whitehead, Weddell, and Groothuis (2016).

instance, with a proportional adjustment factor, the ratio of coefficients for hypothetical and real observations should be the same for all covariates.

While this approach potentially offers some of the same advantages as ours, such as the ability to estimate the effects of treatments that have not been implemented, there are important differences.

First, the data enrichment method requires hypothetical choice variables that are directly comparable to the real choice outcome.[24] For the reasons we discuss elsewhere, this requirement excludes applications such as our analysis of microlending.

Second, for applications satisfying that requirement, the approach does not attempt to capture the motivational factors that potentially cause hypothetical bias to vary across treatment states. For example, Whitehead, Weddell, and Groothuis (2016) use the data enrichment method to estimate a discrete choice model of demand as a function of price in which they include an indicator variable that captures any fixed difference between real and hypothetical choices. Such models do not allow for the possibility that hypothetical bias varies systematically with price, for instance if respondents are less sensitive to social desirability bias at low prices because they don't expect to be judged for accepting a good deal. Our approach can capture such effects because the degree of hypothetical bias implicit in our Step 1 specifications varies with the motivational factors spanned by the vector of hypothetical responses, $\boldsymbol{H}$, which differ across treatment states.

In principle, a model in the spirit of Whitehead, Weddell, and Groothuis (2016) could capture those differences if it also included interactions between the hypothetical indicator and the treatment states. However, one can estimate that model only if the treatments of interest have been implemented. And even then, treatment assignment may be endogenous, in which case a third difference between the methods becomes relevant: when the data-enrichment specification allows hypothetical bias to vary across treatments, endogeneity bias does not vanish as ambient variation in hypotheticals becomes large relative to variation associated with treatment states. Nor does it vanish as hypothetical observations become more numerous.[25] Consequently, endogeneity can bias any estimate of the treatment effect

[24] Some data enrichment methods combine real and hypothetical choice outcomes that are not directly comparable; see Johnston, Ramachandran, and Parsons (2015, ch. 9.2.1.2). While those methods achieve gains in efficiency by exploiting correlations between the error terms, they do not correct for hypothetical bias.

[25] To understand this point, imagine a demand specification $Y_i = \alpha W_i + \beta 1_{\{i \text{ hypothetical}\}} + \gamma W_i \cdot 1_{\{i \text{ hypothetical}\}}$ that includes a treatment indicator, an indicator for hypothetical observations, and an interaction between them. From which coefficient(s) might one infer the treatment effect? The coefficient of the treatment indicator (α) is a natural candidate because it reflects the relationship between real choices and the treatment. However, treatment endogeneity biases it, and no amount of hypothetical data or ambient variation in

derived from the specification even if ambient variation in $\boldsymbol{H}$ is large and hypothetical choice data are plentiful. In contrast, our approach recovers the treatment effect by combining the measured effect of the treatment on $\boldsymbol{H}$, which treatment endogeneity does not impact, with the measured effect of $\boldsymbol{H}$ on choice, which we identify mostly from ambient variation in $\boldsymbol{H}$ when Assumption 3C holds.

Contingent valuation method. Specifications analogous to those used in Step 1 of our method appear elsewhere in the literature on stated preferences, especially in the context of the Contingent Valuation Method (CVM); see e.g., Blackburn, Harrison, and Rutström (1994), Fox et al. (1998), List and Shogren (1998), and List and Shogren (2002). The CVM literature seeks to assess the value of non-market goods, such as those associated with environmental policies, through surveys that pose hypothetical choices. Some CVM studies adjust for hypothetical bias using data from a supplemental survey that elicits both hypothetical and real decisions for a related deliverable good. The analyst performs the adjustment in one of two ways. One approach is to calculate a simple "calibration factor" such as the ratio of hypothetical WTP to real WTP. Another approach is to estimate an individual-level "calibration function" relating real choices to hypothetical responses and characteristics of the respondent.[26] Applying either the calibration factor or the calibration function to hypothetical choices in the main survey yields predictions of real decisions for the non-market good. The calibration-factor approach requires direct comparability between hypothetical responses and real outcomes while the calibration-function approach does not. Even so, studies that use the latter approach typical elicit directly comparable hypothetical responses.

While the goal of the CVM approach is generally to predict a single choice (usually willingness-to-pay) rather than the difference between choices in distinct treatment states, one could in principle use a CVM calibration function to construct treatment effects by taking differences between predicted choices, as in Step 2 of our method. To justify using a CVM calibration function this way, one would need to invoke the Invariant Mapping assumption, as with our method. A critical difference is that our Step 1 specifications

hypothetical responses will cure that problem. The coefficient of the interaction (γ) is not a candidate because it is intended to capture variation in hypothetical bias across treatment states. Likewise, the sum of the coefficients for the treatment indicator and the interaction ($\alpha + \gamma$) is not a candidate because it provides an estimate of the treatment effect on hypothetical choices rather than on real choices. If hypothetical bias can vary across treatment states, the two are not necessarily the same.

[26] Because the typical individual-level calibration specification requires the analyst to elicit a real choice and a hypothetical choice for each respondent, it introduces the possibility of cross-contamination.

implicitly model cross-setting variation in hypothetical bias as a function of motivational attributes spanned by the vector of hypothetical responses $\boldsymbol{H}$, and then use that model to project real choices based on hypothetical responses for alternative treatment states. Our rationale for the Invariant Mapping assumption explicitly references this feature, and indeed we find evidence that the Step 1 mappings are reasonably stable across treatment states in our applications. In contrast, CVM calibration functions do not exploit cross-setting variation in motivational attributes that potentially affect hypothetical bias. Indeed, when estimating a calibration function, these studies typically use data from a single setting, and examine different settings only to check stability.[27] In practice, studies have found that calibration mappings vary across deliverable goods and are sensitive to the context of choice.[28] This instability calls the validity of the Invariant Mapping assumption into question for the class of CVM specifications used in the literature. The apparent failure of this assumption is unsurprising because hypothetical bias likely depends on a variety of motivational factors that no single hypothetical choice variable adequately spans.

Additionally, even if there were a stable structural relationship at the individual-level between Y and the individual's motivations, the empirical relationship between Y and a noisy measure of motivation such as Y^H at the individual-level may vary across contexts due to the attenuation bias, especially if the means of Y differ. Our Step 1 specifications also address this source of instability in two ways. First, aggregation over individuals within each setting mitigates the effects of noise in hypothetical responses. Second, the analyst can use the IV technique we mentioned in our discussion of measurement error to address any remaining noise.

Meta analysis. Our Step 1 specifications are more closely related to those used in meta-analyses relating variation in hypothetical bias across studies to study characteristics, S_j (where j indexes studies); see, e.g., List and Gallet (2001), Little and Berrens (2004), Murphy et al. (2005), Penn and Hu (2018), as well as Nelson (2015), Rolfe, Brouwer, and Johnston (2015), and Boyle, Kaul, and Parmeter (2015) in the benefit transfer literature.

[27] List and Shogren (2002) estimate a single calibration function pooling WTP data for multiple goods, but likewise are mainly concerned with testing stability of the calibration across goods. Because they do not estimate a model linking the variation in hypothetical bias across goods to variation in motivational attributes, their specifications do not allow the analyst to predict differences in bias across treatment states.

[28] For example, according to Fox et al. (1998), "calibration seems commodity specific, and as such, CVM-X should proceed on a case-by-case basis until enough evidence reveals a systematic pattern leading to a general calibration function." List and Shogren (1998) likewise conclude that "calibration is good- and context-specific." See also Blackburn, Harrison, and Rutström (1994), List and Shogren (2002), Ajzen, Brown, and Carvajal (2004), and Johansson-Stenman and Svedsäter (2012).

Typically, these characteristics include coarse features of the choice problems such as whether they involve public or private goods, as well as controls for the characteristics of respondent populations and the protocols used to elicit hypothetical choices.

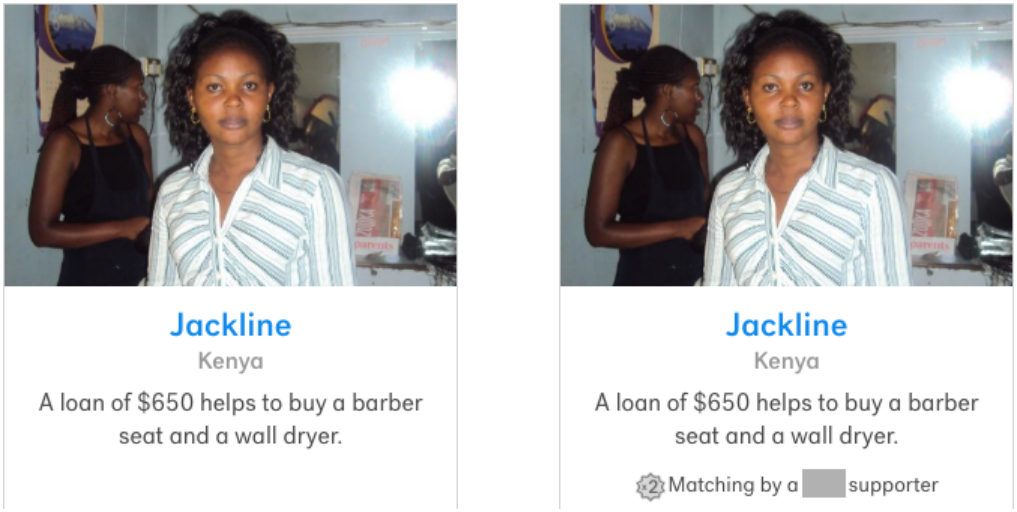
There are three central differences between meta-analysis specifications and those used in our Step 1. First, by focusing on hypothetical bias, meta-analyses exclude applications, such as our analysis of microlending, in which real and hypothetical responses are not directly comparable. Second, meta-analysis specifications do not control for hypothetical responses that capture variation in the motivational factors behind hypothetical bias. Consequently, if one used them to impute hypothetical bias for the same setting in two different treatment states, the answers would be identical. Third, in our method, controls for elicitation protocols and respondent populations are not needed in Step 1 as long as the analyst holds the set of hypothetical evaluations and the protocols used to generate them fixed across settings (the preferred procedure). For applications in which the real-choice population is the same across settings (such as those explored in Sections 3 and 4), the same is true for respondent populations.

Other literature. Another related branch of the literature on stated preferences attempts to correct hypothetical bias by "fixing" the elicitation protocol (e.g., Cummings and Taylor 1999, Jacquemet et al. 2013, and Blamey, Bennett, and Morrison 1999, and Loomis, Traynor, and Brown 1999) or appropriately screening responses (Champ et al. 1997 and Whitehead, Weddell, and Groothuis 2016). For our method, there is no need to assume that any protocol or screening procedure yields an unbiased prediction. Instead, each produces a potentially useful predictor.

In Appendix D, we discuss other related work in marketing (Infosino 1986; Jamieson and Bass 1989; Morwitz, Steckel, and Gupta 2007), political science (Jackman 1999; Katz and Katz 2010), and neuroeconomics (Smith et al. 2014). We also describe connections to linear factor models, synthetic controls, and statistical surrogates.

3 Application: Microfinance Contributions

Next we turn to a field application. To boost fundraising, non-profit organizations often inform potential contributors that other donors have agreed to match contributions (List 2011). How well does this strategy work? In this section, we use our method to determine the impact of matching provisions in the context of microfinance. This application shows



(a) Unmatched profile (b) Matched profile

Figure 2: Loan Profiles with Matching Indicator

that, by combining hypothetical responses with real choice data, our method accurately recovers treatment effects in a field setting where the real choice data alone are insufficient due to endogenous treatment selection and the absence of good instruments.

We focus on a large microfinance crowdsourcing website, which displays profiles of potential borrowers and allows website visitors to contribute to their loans. Contributors are typically socially minded individuals in developed countries; borrowers are typically developing country residents who request funds for various projects (such as business, agricultural, home, or health expenses). The website allows sponsors to match contributions, dollar for dollar, to loan profiles meeting criteria the sponsors specify. When a loan is eligible for a match, the profile displays an indicator (as shown on Figure 2).

In this application, the unit of observation j is a loan profile and the potential outcome $Y_j(w)$ is fundraising velocity for the first 24 hours after the loan appears on the website. We transformed velocity using the inverse hyperbolic sine to reduce the impact of outliers.[29] The treatment $W_j \in \{0, 1\}$ specifies whether the loan is matched by sponsors, who may be individuals or collectives such as churches or community groups. The treatment effect of interest is the impact of matching on fundraising velocity. Significantly, sponsors often specify criteria for loan selection (for example, based on the borrower's gender, region, sector, loan size, risk, and/or number of days until expiration). As a result, although real

[29]We define fundraising velocity as the number of (non-matching) dollars raised per day. For loans that fully fund in less than 24 hours, we calculate velocity based on the funding period. The inverse hyperbolic sine resembles the natural logarithm but is defined at zero.

choice data are available, the potential endogeneity of matching complicates estimation. Endogeneity arises from correlations between the preferences of sponsors and contributors. Because it is difficult to imagine characteristics of postings that would affect the choices of sponsors but not of contributors, instruments are not readily available.

We compare estimates from observational data using standard methods and our method to a "ground truth" estimate based on a field experiment in which we introduce randomly assigned matches (similar to Karlan and List 2007; Huck and Rasul 2011).[30]

3.1 Data

In this section, we describe the observational data and survey data on hypothetical responses used in our analysis, as well as the experimental data we use to establish ground truth.

3.1.1 Observational data

Through a collaboration with the website, we observe 11,668 loan profiles for borrowers seeking $1,000 or less posted between October 14, 2019, and November 3, 2019 (omitting a random subsample that served as the treatment group for our experiment, as described below). We retain 9,623 profiles (82%) that were either unambiguously matched (matched for at least 90% of the first 24 hours after their initial posting) or unmatched (matched for no more than 3% of the first 24 hours). After dropping the remaining 18% of profiles, which were matched for intermediate fractions of the first 24 hours, we create a binary treatment indicator.[31] According to this indicator, 623 (6.5%) of the retained profiles were matched, so the treatment is rare. For each profile, our data include descriptive characteristics, whether it was matched, and how quickly it raised funds.

3.1.2 Hypothetical responses

Separately, we collected responses to hypothetical questions concerning a subset of the loan profiles from 833 participants recruited through Amazon Mechanical Turk. We selected 200 unmatched and 100 matched loan profiles at random from the observational sample for our exercise, oversampling matched loans. For each participant, we selected 30 of these profiles at random, 20 drawn from the sample that was not matched on the website, and

[30] This experiment was preregistered (AEARCTR-0004885). Deviations from the plan are described in Appendix E.1.2.

[31] Observational methods yield similar estimates of the treatment effect when we retain all profiles and regress the outcome on the fraction of time each profile was matched during the first 24 hours.

10 from the sample that was matched. Participants initially viewed an overview page with a large collection of "thumbnail" profiles that reflected the overall prevalence of matches among active loans on the website. Then participants saw each of their selected 30 profiles either in the same treatment state as on the website or edited to appear in the opposite state (with the matching funds indicator either added or removed). We displayed 15 of the 30 selected profiles as unmatched (10 of which were actually unmatched on the website) and 15 as matched (5 of which were actually matched on the website).

Participants provided four hypothetical evaluations for each (real or counterfactual) loan profile. They (1) predicted which quintile of fundraising velocity each profile would attain, and indicated the likelihood that (2) they or (3) a typical user would lend $25 to it (both 7-point scales, from very unlikely to very likely). We incentivized the first question: respondents who predicted the correct quintile for a randomly chosen profile (among those displayed exactly as they appeared on the website) received a bonus of $2. After participants rated all 30 profiles, we (4) posed the following task: "Suppose you have decided to make a total of ten $25 loans to postings among the 30 you just viewed. Which 10 would you pick?" Through this process, we generated on average slightly more than 40 sets of evaluations for each matched or unmatched loan profile (minimum 39, maximum 46). The survey included several features that encourage participants to submit thoughtful responses, as detailed in Appendix E.2.

Importantly, none of the four hypothetical evaluations is directly comparable to the choice outcome (funding velocity). Consequently, one cannot assess hypothetical bias in this application by comparing the real outcome and the hypothetical assessments.[32]

3.1.3 Ground truth experiment

We established ground truth through an experiment. Starting on October 27, 2019, we assigned all new loan listings for borrowers seeking $1,000 or less either to a treatment group (roughly 10%) or a control group (roughly 90%).[33] We established a sponsorship account for loans in the treatment group and used it to ensure that contributions to them were matched for the first 24 hours after they appeared on the website. We stopped adding loans to our sample once the funds in the sponsorship account were depleted. The resulting

[32]The analyst could, in principle, ask respondents to predict *absolute* (rather than relative) funding velocity, which would be directly comparable to the choice outcome. However, that measure depends on factors respondents are poorly equipped to assess such as the rate at which visitors come to the website and view the pertinent posting. For such measures, the Invariant Mapping assumption is therefore likely to fail.

[33]The treatment group includes loans with identifiers ending in zero, and the control group includes loans with identifiers ending in any other number.

treatment group includes 109 loans, and the resulting control group includes 982 loans.

Other sponsors continued to match loans during the course of our experiment. For the treatment group, the website used matching funds from our sponsorship account only if the loan did not meet the criteria set for any other active sponsorship account. Loans that would be matched irrespective of our intervention are always-takers; they correspond to matched loans in the observational sample. Loans that would not have been matched in the absence of our intervention are compliers.[34]

3.2 Treatment effects

Table 1 contains estimated treatment effects for matching provisions derived through a variety of methods. Because we ran an experiment, we can estimate a ground truth measure of the treatment effect (Column (1)). In the experiment, the intention to match is random, so we use it as an instrument for the endogenous matching indicator. The standard instrumental variables (IV) estimate of the LATE is 1.24 (s.e. 0.33), which measures the effect on compliers (unmatched loans). This coefficient implies that matching increases funding per day by roughly 245% on average.

Endogeneity bias in observational data. Next we attempt to recover treatment effects using only the observational data. It is difficult to unravel the structure of the process that renders matching provisions endogenous: conditioning the analysis on sponsors' potential matching criteria is impractical and potentially ineffective,[35] and good instruments are not readily available. Because the types of loan profiles that draw matching funds also tend to attract contributions, estimators that do not address this endogeneity exhibit substantial bias. The simple difference in means implies an estimated treatment effect of 2.55 (Column (2)), more than twice ground truth. Controlling for standard covariates such as the gender and industry of the loan profile does not reduce this bias, regardless of whether we insert each factor linearly (Column (3)) or flexibly control for linear and interaction terms using ARB (Column (4)) as in Athey, Imbens, and Wager (2018).[36] We reject equality between each of

[34] Because we always carried out our intention to match contributions for loans in the treatment group, our design rules out the existence of never-takers and defiers.

[35] Even in a high-dimensional model that controls for all the criteria sponsors can specify, the treatment is likely endogenous. Sponsors may decide to match certain types of proposals based on transient factors that may also influence contributions, such as the attractiveness of postings within particular categories at the time of the matching decision. Conditional on time and criteria, there is no variation in treatment, which means standard approaches are infeasible.

[36] These regressions include interactions with the covariates recentered at their means within the control (unmatched loans). Consequently, if endogeneity were not an issue, the coefficient of the match indicator

Table 1: Estimated Treatment Effects from Microfinance Application

	Ground Truth	Observational Methods			Our Method: Hypotheticals as Predictors						
	Experiment (IV)	Diff	OLS	ARB	Low dimensional				High dimensional		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Estimated effect of matching	1.24	2.55	3.21	3.01	0.62	0.69	1.48	1.28	1.80	1.00	1.38
	(0.33)	(0.33)	(0.31)	(0.29)	(0.32)	(0.31)	(0.36)	(0.38)	(0.25)	(0.18)	(0.26)
	[0.32]	[0.34]	[0.30]	[0.29]	[0.33]	[0.30]	[0.39]	[0.39]	[0.39]	[0.31]	[0.47]
Test: = ground truth, p-value	1	0.01	0.00	0.00	0.18	0.21	0.63	0.93	0.27	0.58	0.80
(adjusted)	(1)	(0.02)	(0.00)	(0.00)	(0.46)	(0.52)	(0.88)	(0.93)	(0.55)	(0.90)	(0.95)
Controls			X	X		X		X	X	X	X
Hypotheticals:											
. . . avg. hypothetical eval.					X	X	X	X	X	X	X
. . . freq. hypothetical eval.										X	X
. . . 2nd order							X	X	X		X
interactions				X					X		X
Sample size	1091	300	300	300	300	300	300	300	300	300	300
Observed matched	use randomized variation	endogenous			endogenous						
Observed unmatched	use randomized variation	endogenous			endogenous						

Estimates of the effect of matching on the inverse hyperbolic sine of fundraising velocity, within the first day. Controls include dummies for gender, region, and sector. 'Avg. hypothetical eval.' includes the mean responses concerning projected quintile for fundraising velocity, contribution likelihoods (respondent and typical user), and funding allocation. 'Freq. hypothetical eval.' includes the frequency of "at least" each potential response to each hypothetical question (for instance, the frequency of respondents projecting the second or higher quintile, the third or higher quintile, etc.). '2nd order' includes quadratic terms for the mean responses and frequencies of each hypothetical response (if used). 'interactions' includes all two-way interactions between mean responses, frequencies of each hypothetical response (if used), and the controls. Analytical standard errors in parentheses, bootstrap standard errors in square brackets. Adjusted p-values apply a stepdown procedure to adjust for multiple hypothesis testing for the tests in this table (Romano and Wolf 2010; List, Shaikh, and Vayalinkal 2023).

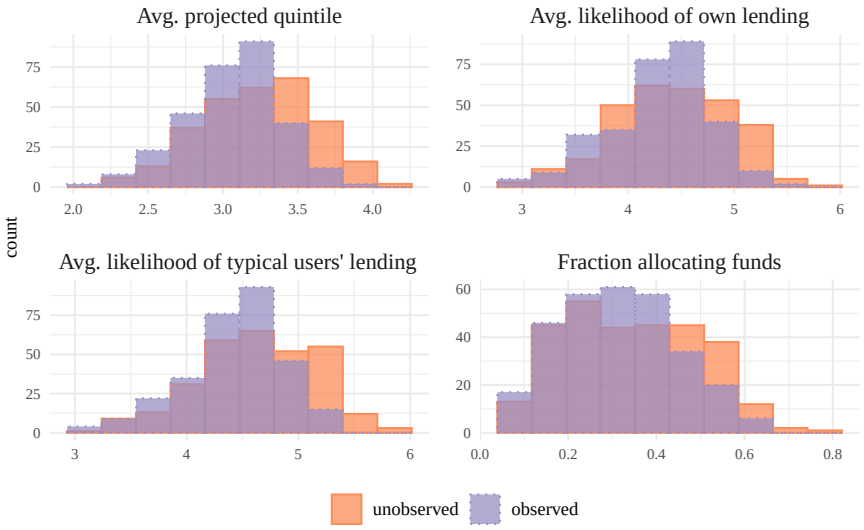


Figure 3: Overlap in hypothetical evaluations for loan × treatment states that are observed (blue) vs. unobserved (red) in the data.

these estimates and the ground truth.

Estimates based on the two-step method. Next we turn to estimates based on hypothetical evaluations. Because none of the four hypothetical choice variables are directly comparable to the real choice outcome (velocity), the simple difference-in-hypotheticals estimator discussed in Section 2.2.1, $\hat{\tau}_{\text{hyp}}^{simple}$, is inapplicable. While one could still try to mimic that estimator,[37] such alternatives perform relatively poorly; see Appendix E.3.1. Consequently, this application provides a good opportunity to see how our estimators perform not only when treatment endogeneity poses a serious challenge, but also when simpler hypothetical choice methods miss the mark.

Before applying the two-step procedure, an analyst would perform two types of diagnostic checks. The first diagnostic evaluates overlap, as suggested in Section 2.3.1. Figure 3 shows that, for most of the evaluations of profiles in counterfactual states (red), there are indeed loans with similar evaluations in their assigned states (blue). Consequently, our method requires only modest extrapolation (for profiles and treatments that are highly desirable).

would measure the average treatment effect for the control (ATC). Using this specification ensures that the procedure targets the same object as the experiment, i.e., the LATE (which coincides with the ATC).

[37] For example, by picking a threshold (such as 6 on a 7-point scale), we can convert any hypothetical response into a dichotomous "lending decision," compute the implied "lending frequencies" for each setting and treatment state, take the percentage difference, and average over settings. The result is potentially interpretable as an estimate of the proportional effect of matching on loan volume per day.

The second type of diagnostic speaks to the validity of Assumption 3C. The vast majority of variation in hypotheticals is ambient (condition (i)): treatment accounts for only 4-7% of their variation as measured by the ratio $\frac{\text{var}(W_j(H_j(1)-H_j(0)))}{\text{var}(H_j(\tilde{w}))}$ for the regressors used in Column (8) of Table 1. Hypotheticals also have decent explanatory power (condition (iii)): a regression of $Y_j(0)$ on $\boldsymbol{H}_j(0)$ and $\boldsymbol{X}_j$ among the untreated observations yields $R^2 = 0.21$. These diagnostics suggest that the asymptotic bias of the two-step estimator $\hat{\tau}$ is small per Proposition 2.

We use the two-step procedure to recover the average treatment effect. In the first step, we estimate the mapping from hypothetical evaluations $\boldsymbol{H}_j(W_j)$ to choices Y_j using data on the matched and unmatched loan profiles for which we gathered hypothetical responses. Each specification includes the mean response for each of the four hypothetical evaluations (projected quintile, likelihood of own lending, likelihood of typical user lending, and funding allocation); the specifications differ according to whether they include other regressors. We reweight the sample to ensure that it accurately reflects the mix of loans on the website. In the second step, we use the estimated mapping from the first step to predict outcomes for both states and take the difference. We then compute the average of the difference for unmatched loans. With this sample restriction, the procedure targets the same object as the experiment, i.e., the LATE (which coincides with the average treatment effect for the control, ATC).

Table 1 exhibits a low-dimensional specification that includes the average of each hypothetical evaluation linearly (Column (5)), as well as ones that add standard controls (Column (6)), squared hypothetical evaluations (Column (7)), and both (Column (8)).[38] It also includes high-dimensional specifications based on our ARB estimator that add interaction terms (Column (9)), distribution detail for each possible hypothetical response (Column (10)), and both (Column (11)). Estimates range from 0.62 to 1.80, with richer specifications generally producing results closer to ground truth. Statistical tests fail to reject the hypothesis that each of the estimates using our method coincides with the ground truth. But, crucially, our method requires only a *hypothetical* experiment.

The procedure performs less well when Step 1 employs a specification relating real choices to only one of the four hypothetical choice measures. This finding is expected because a single variable is less likely to span the pertinent motivations. For univariate specifications in which the hypothetical variable enters linearly, estimates of τ range from -0.02 to 0.29. The addition of higher order terms moves the estimates toward ground truth

[38] Quadratic terms may be particularly useful in this context because, in contrast to the snack application, the real and hypothetical choices are measured in different units.

for the quintile projection but achieves only modest improvement for the three alternatives. Without knowing ground truth, the best choice among the four measures would be unclear. See Appendix E.3.2 for details.

Estimated mappings. As a side product, the first step of our method yields estimates of the mapping between hypothetical evaluations and real choices, $\mu(\boldsymbol{h}, \boldsymbol{x})$. We report the estimated relationships corresponding to Columns (5-8) in Appendix Table A1. Multiple types of hypothetical responses contribute to the predicted outcomes. When controlling for respondents' projections of their own lending decisions and those of others like them, the projections about others are the ones that robustly exhibit a strong and directionally intuitive relationship to funding velocity. This result corroborates a finding of Rothschild and Wolfers (2011), that vicarious questions can generate usefully predictive information.[39]

Estimates based on the sample selection and bounding methods. In Section 2.3.2, we proposed two procedures for addressing endogeneity without invoking Assumption 3C. These approaches corroborate the results obtained through our basic two-step estimator. Applying our sample selection correction to the same specification as in Column (8) of Table 1, we obtain a treatment effect estimate of 1.31 (bootstrap s.e. 0.46).[40] One can also bound the treatment effect without making any assumption about the treatment assignment process. When imposing only "natural" bounds on potential outcomes in the spirit of Manski (1990)—specifically, that funding rates are non-negative and loans are never fully funded in less than one minute—one can only say that the average treatment effect falls within the interval $[-5.02, 9.85]$.[41] Adding mapping invariance and linearity substantially narrows these bounds to $[0.37, 1.88]$, based on the specification in Column (8) of Table 1. Adding a constraint that matching cannot *reduce* funding for any loan shrinks the bounds further, to $[0.67, 1.25]$.[42]

[39] While the same pattern appears in the estimated mappings for the high-dimensional specifications (Table 1 Columns (9-11)), we do not report those mappings because coefficients in the ARB procedure do not have straightforward interpretations. As discussed in Section 2.4, one cannot reinterpret the Step 1 specifications as relationships between hypothetical bias and the settings' subjective and objective characteristics, $\boldsymbol{H}(w)$ and $\boldsymbol{X}$, because the hypothetical variables are not directly comparable to the choice outcome.

[40] Estimates corresponding to Columns (4)–(7) of Table 1 are within about 0.3 of the estimates without correction, see Appendix Table A2.

[41] Formally, the "natural" bounds are $\underline{Y} = 0$ and $\overline{Y} = 14.87$. Among observed postings, funding velocity never exceeds the upper bound: $\max_j Y_j = 11.87$, which is equivalent to full funding in 20 minutes for the largest loans ($1,000).

[42] Note that the point estimate based on the same assumptions can be outside the bounds due to sampling uncertainty. It is challenging to compute confidence intervals because the number of constraints grows with the sample size; we describe an approximation in Appendix C.6.3.

Table 2: Heterogeneity by Compliance Group in the Microfinance Application

	Experiment	Our Method		Proportion
	IV (1)	Low Dimensional (5)	High Dimensional (11)	of Observational Sample
Estimated effect of matching				
..... on compliers (LATE/ATC)	1.24 (se 0.32)	0.62 (se 0.33)	1.38 (se 0.47)	93.5%
..... on always-takers (ATT)	cannot be estimated	0.11 (se 0.19)	0.75 (se 0.38)	6.5%
... average (ATE)	cannot be estimated	0.59 (se 0.32)	1.34 (se 0.44)	100%
Test: equal effects, p-value	—	0.03	0.24	
(adjusted)		0.05	0.24	

The first row of estimates reproduces results from Table 1, columns (1), (5), and (11) (as indicated in the column headings). Standard errors in parentheses are based on the bootstrap. Adjusted p-values apply a stepdown procedure to adjust for multiple hypothesis testing for the tests in this table (Romano and Wolf 2010; List, Shaikh, and Vayalinkal 2023).

Heterogeneity in treatment effects By applying the second step of our method to designated observations, one can estimate average treatment effects for any desired subset of settings. We illustrate this capability in Table 2. The first row reproduces selected estimates from Table 1, including the IV estimate, as well as two measures obtained through our method (corresponding to the low and high dimensional specifications in, respectively, Columns (5) and (11)). As noted earlier, our method approximates ground truth for the LATE/ATC. The ATC, or effect on compliers, is relevant for some purposes but not for others. For example, to evaluate the impact of eliminating the website's matching provisions, one would need to determine the effect of matching on funding velocity for loans that are currently match-eligible (always-takers). Our method implies that the average treatment effect on the treated (ATT) is less than half the LATE/ATC for both specifications, as shown in the table's second row. Alternatively, to evaluate the effect of introducing a match and making it universal, one would need to determine its overall effect (ATE). The estimated ATEs appear in the third row, and are close to the LATE/ATCs because the population of always-takers is relatively small (6.5% of the total). The corresponding cells for the instrumental variables estimator are empty because the experiment cannot reveal either of these effects. The magnitude of the ATT relative to the LATE/ATC has an immediate policy implication: the microfinance platform may be able to raise more funds by inducing sponsors to shift their support from the currently matched loans to the currently unmatched loans.

4 Application: Snack Demand

Next, we evaluate the performance of our approach in the laboratory. We use data from an experiment to simulate applications where we observe choices for each setting in a single treatment state. The main advantage of the experiment is that we learn what would actually happen in the other treatment state. This feature allows us to (1) test the Invariant Mapping assumption directly by estimating $\mu_w(\boldsymbol{h}, \boldsymbol{x})$ separately for each treatment state w, (2) simulate a variety of treatment selection processes, including those with no treatment variation, random treatment assignment, and endogeneity, and (3) assess how well the method recovers heterogeneous treatment effects.

The laboratory experiment simulates applications in which the objective is to estimate price sensitivities for a collection of related goods. The goods we study are familiar snack foods. We chose this domain because we expect hypothetical bias to be important and to vary across items. For example, people may downplay their preferences for unhealthy or unpopular snacks and exaggerate their preferences for healthy or popular alternatives. Treatment states $w \in \{0, 1\}$ correspond to prices of \$0.25 or \$0.75, respectively. $Y_j(w)$ denotes aggregate demand for good j at the price corresponding to w. The treatment effect of interest is either the average price response $\frac{1}{J}\sum_{j=1}^{J}[Y_j(1) - Y_j(0)]$, or the responses for individual goods.

We apply our method to datasets containing one real observation for each good (demand at a single price). We extract those samples from the full dataset, which contains two real observations for each good (demand at both prices). We use the full dataset to measure true price responses (ground truth).

4.1 Procedures and data

Each of 365 subjects was assigned to one of several groups, described below.[43] Subjects were told that their sessions consisted of two stages. The first involved a computer-based choice or rating task lasting roughly 30 minutes. The second was a 30-minute waiting period. Subjects were asked not to eat anything during the waiting period unless a snack was provided (according to the rules).

[43] We conducted the experiment at the Stanford Economic Research Laboratory (SERL) between January, 2011, and October, 2012. We gathered most of the data between January and May 2011, adding data on norms and hypothetical willingness to pay in November 2011 and data on vicarious assessments in October 2012. Stanford University's IRB reviewed and approved the protocol. The participation fee ranged from \$20 to \$30. We adjusted the fee upward when the response rate to our subject solicitation was low and downward when it was high. Sessions took place in mid-afternoon when subjects are typically hungry.

In the first stage of each session, a group of subjects decided whether to purchase each of $J = 189$ snacks at a given price, \$0.25 or \$0.75. For one subgroup, these decisions were real and provide the basis for measuring $Y_j(w)$. For a second subgroup, they were hypothetical. Other groups were asked to rate the same snacks according to various subjective criteria, with price a factor in some questions. Together, these hypothetical responses provide the basis for measuring $\boldsymbol{H}_j(w)$.

The stimuli (food items or item-price pairs) were presented in random order. Most groups consisted of roughly 30 subjects. For experiment details and a complete catalog of the groups along with sample sizes and a screenshot for a representative question, see Appendix F.1 and Figure A1.

4.1.1 Real choices

The subjects who made real choices were informed that we would select one decision at random and implement it during the 30-minute waiting period.[44] Although the chance of implementing any given choice was low, differences between real and hypothetical purchase frequencies were substantial, and in the expected direction, as discussed in Section 4.1.3.[45]

In observational data, we might observe such demand at a single price for each good, possibly without variation or set endogenously. Our design allows us to observe demand at both prices, which we use to establish ground truth. We then mimic observational data by restricting the estimation sample to observations of demand at a single price for each good.

4.1.2 Hypothetical evaluations

Other participants provided various hypothetical evaluations, designed to span underlying motivations as well as factors that cause hypothetical choices to diverge from real ones. We gathered different evaluations from different respondents to minimize interference and

[44] By construction, it follows that the demand for each snack item does not depend on the prices of the other items. Our framework could accommodate substitution across products by specifying the price of every good in each hypothetical question.

[45] Real purchase frequencies were not significantly different in a group of participants whose odds of implementation were one in 5 decisions rather than one in 378. See Appendix F.2 for details. It is not surprising that participants in the "real choice" group viewed their choices as real: they had as much at stake as someone making a single purchase decision (because they knew we would definitely implement one choice), and taking the task less seriously did not reduce the subject's time commitment. Notably, similar conclusions were reached by Carson, Groves, and List (2014) based on theoretical principles and experimental evidence, and by Kang et al. (2011) based on fMRI data. Consistent with these findings, a survey paper by Brandts and Charness (2009) found no support for the hypothesis that differences between the strategy method and the direct response method increase with the number of contingent choices.

fatigue.

Several groups made hypothetical choices. The literature on stated preferences explores a variety of protocols for eliciting such choices. An advantage of our approach is that it is not necessary to select only one elicitation protocol. We employed multiple protocols, each with a separate group. The "standard" protocol mimicked the real choice protocol, except that no choice was implemented. A second protocol employed a "cheap talk" script (as in Cummings and Taylor 1999) that encouraged subjects to take the hypothetical choices seriously,[46] a third elicited likelihoods rather than Yes/No responses (analogously to Champ et al. 1997), a fourth asked about the likely choices of same-gender peers (to eliminate image concerns and thereby potentially obtain more honest answers, analogously to Rothschild and Wolfers 2011), and a fifth elicited hypothetical willingness-to-pay (WTP) rather than Yes/No responses.

Some groups provided subjective ratings. Depending on the group, subjects reported their anticipated degree of happiness with each potential purchase, the anticipated degree of social approval or disapproval for each potential purchase, how much they liked each item, evaluations of regret, measures of temptation, expected enjoyment (ignoring considerations of diet or health), perceptions of health benefits, impact of consumption on social image, and the perceived inclination to overstate or understate the likelihood of a purchase.

4.1.3 Patterns of real and hypothetical choices

In the real-choice (ground truth) group, on average, 28% of people elect to purchase the average snack when the price is \$0.25. When the price rises to \$0.75, the average purchase frequency declines by 7.5 percentage points ($\tau = -0.075$, standard error 0.004). This response varies across snacks: its standard deviation is 6 percentage points.

On average, hypothetical choices overstate purchase frequencies by 28% (7 percentage points). They also overstate how sensitive choice is to prices by a factor of more than two-to-one ($\tau = -0.159$, standard error 0.006): see Figure 4a. The variance of hypothetical choice is also more than twice that of hypothetical choice (see Appendix F.3). These findings suggest that, in our data, people may generally exaggerate the sensitivity of their choices to variation in conditions, contrary to Lusk and Schroeder (2004).

In our data, alternative hypothetical choice protocols do not systematically 'fix' hypothetical bias. The correlation between real demand and the standard-protocol hypothetical

[46] We thank Laura Taylor for generously reviewing and suggesting changes to the script, so that it would conform in both substance and spirit with the procedure in Cummings and Taylor (1999).

demand is higher than for any alternative protocol, which casts doubt on the hypothesis that any of the alternative protocols improve the informational content of hypothetical choices. Also, unless one makes a fortuitous assumption about the mapping from Likert scales to choices, the standard protocol exhibits less bias on average.[47]

Although hypothetical choices are poor predict*ions* of real choices, they are strongly correlated with real choices, and consequently may be useful predict*ors*, particularly when used in combination (assuming they elicit different and potentially complementary information). Critically for our method, real choices and hypothetical responses are generally consistent with the Invariant Mapping assumption. Figure 4b illustrates this point for standard hypothetical choices. Each data point represents the real and hypothetical purchase frequencies for a given item at a specified price—$0.25 for the orange squares and $0.75 for the purple dots. The figure also shows best-fit lines for each price. In addition to being systematic, these fitted relationships are also virtually identical, and we do not reject equality (p-value 0.50). The relationships between real choices and hypothetical responses are similarly stable for most other "first-step" specifications as Invariant Mapping requires.[48]

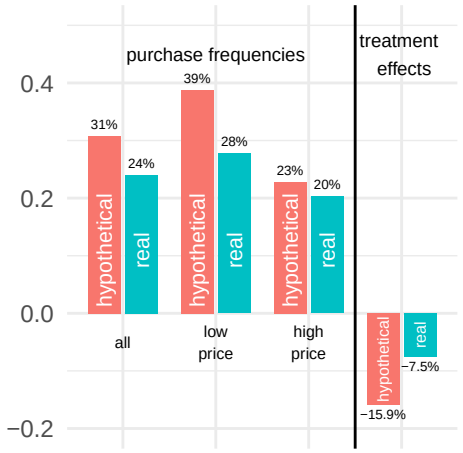
4.2 Treatment effect estimates with problematic treatment assignment

In contrast to estimators based solely on observed choices, our method can recover treatment effects in applications for which there is no real-world variation in the treatment of interest, or only endogenous variation.

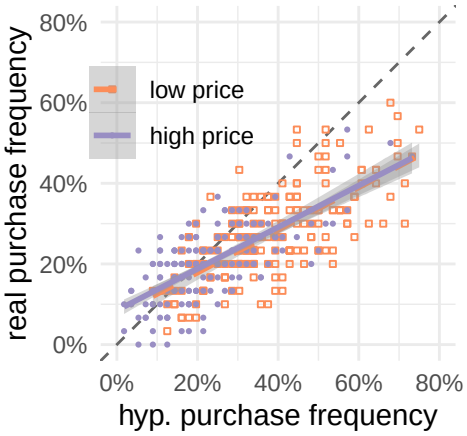
For either type of environment, the analyst would first assess overlap of the hypothetical variables between treatment states, as suggested in Section 2.3.1. Appendix F.4 shows that, for an analysis with all snacks at the low price (or all snacks at the high price), overlap is almost complete for all hypothetical evaluations except the WTP purchase frequency. Because the distribution of that variable at the low price does not span much of the corresponding distribution at the high price, using it would require substantial extrapolation. We therefore exclude it from our multivariate specifications throughout.

[47] For both points, see Appendix Table A4, panel (a). Concerning the second point, hypothetical choices elicited with the cheap talk script or using WTP measures are further from ground truth. Counting the highest likelihood of one's own purchase ("very likely to purchase") as a choice yields a modest improvement, and the equivalent variable derived from vicarious likelihood performs even better, but both perform worse at other dichotomization thresholds.

[48] See Appendix Tables A5 and A6. The main exceptions are specifications that include a dichotomized WTP choice frequency, which fails the overlap diagnostic. Results based on nonparametric regression are also similar; see Appendix Figure A2.



(a) Choices vs. Price



(b) Real vs. Hypothetical Choices

Figure 4: Real and Hypothetical Choices

Panel (a) plots the average proportion purchasing by price. Hypothetical choices are biased. However, as shown in panel (b), there is a predictable relationship between real and hypothetical choices. Panel (b) plots item-price pairs. Separate regression lines for the $0.25 choices and the $0.75 choices are shown with error bands.

4.2.1 Unseen treatments

In this section, we use our experimental data to simulate an environment in which the analyst observes all snack purchases at a single price, and then uses the two-step procedure to estimate the average treatment effect. In the first step, we estimate the mapping from hypothetical evaluations to purchase frequencies. In the second step, we use the estimated mapping to predict purchase frequencies for each snack at both prices, compute the average predicted frequency at each price, and take the difference.

Table 3 reports results. Column (1) reproduces the true average treatment effect ($\tau = -0.075$). Columns (2)–(6) use our two-step method; each employs a different specification of the mapping from hypothetical evaluations to choices (Step 1). Different specifications include different combinations of hypothetical responses, physical characteristics (grams per serving and amount of seven nutrients), and higher order terms. All specifications include simple versions of the hypothetical variables (averages and dichotomous frequencies); some add detailed information on the full distributions of Likert-scale responses.[49] We estimate the prediction model using OLS for the first two specifications and ARB for three

[49]As noted above, we omit the WTP variable because it fails the overlap diagnostic. Including the WTP variable yields similar estimates when all snacks are observed at the high price, but causes most estimates to be farther from ground truth when all snacks are observed at the low price; see Appendix Table A3. The reason is that, due to the failure of overlap, predicting demand at the high price requires more extrapolation.

Table 3: Estimating Treatment Effects without Variation in Treatment (multivariate)

	Ground Truth	Our Method: Hypotheticals as Predictors				
		Low Dimensional		High Dimensional		
	(1)	(2)	(3)	(4)	(5)	(6)
Observing all snacks at high price						
Estimated effect of high price	-0.075	-0.082	-0.075	-0.073	-0.084	-0.077
	(0.004)	(0.010)	(0.010)	(0.012)	(0.005)	(0.012)
	[0.004]	[0.009]	[0.010]	[0.017]	[0.016]	[0.024]
Observed at high price	All			All		
Observed at low price	All			None		
Test: = ground truth, p-value	1	0.50	0.99	0.92	0.60	0.93
(adjusted)	(1)	(0.94)	(0.99)	(1.00)	(0.95)	(1.00)
Observing all snacks at low price						
Estimated effect of high price	-0.075	-0.100	-0.096	-0.098	-0.094	-0.096
	(0.004)	(0.008)	(0.008)	(0.009)	(0.004)	(0.013)
	[0.004]	[0.007]	[0.007]	[0.015]	[0.014]	[0.017]
Observed at high price	All			None		
Observed at low price	All			All		
Test: = ground truth, p-value	1	0.00	0.00	0.12	0.18	0.19
(adjusted)	(1)	(0.01)	(0.03)	(0.58)	(0.67)	(0.63)
Controls			X	X	X	X
Hypotheticals:						
. . . all hypothetical choices (excl. WTP)		X	X	X	X	X
. . . detailed hypothetical eval. (excl. WTP)					X	X
2nd order + interactions				X		X
Sample size (outcome)	189 ($\times 2$)	189	189	189	189	189

Estimates of the effect of the high price (vs. low price) on the real purchase frequency. Analytical standard errors are in parentheses; bootstrap standard errors in square brackets are based on 1,001 bootstrap samples. Analytical standard errors for the high-dimensional specifications treat average evaluations as fixed, whereas all other standard errors consider estimation error in the average evaluations across snacks. Adjusted p-values apply a stepdown procedure to adjust for multiple hypothesis testing for the tests in all panels of this table (Romano and Wolf 2010; List, Shaikh, and Vayalinkal 2023). For results including WTP see Appendix Table A3.

high-dimensional specifications. The estimates of τ are close to the true average effect when all snacks are observed at the high price (between -0.073 and -0.084). Accuracy is somewhat lower when all snacks are observed at the low price (between -0.094 and -0.100).

For this application, the procedure also performs reasonably well when Step 1 employs a univariate specification relating real choices to a single hypothetical choice variable. Focusing on the variables for which overlap is adequate, the estimates of τ range from -0.50 to -0.86 when all snacks are observed at the high price, and from -0.48 to -0.101 when all snacks are observed at the low price; see Appendix Table A4 panel (b).[50] Thus, while specifications employing multiple hypothetical variables yield somewhat greater accuracy and stability than those employing a single hypothetical variable, the difference is not as large as for our field application.

Estimated mappings. Once again, as a side product, the first step of our method yields estimates of the mapping between hypothetical evaluations and real choices, $\hat{\mu}_w(\boldsymbol{h}, \boldsymbol{x})$. These estimates show that variation in real choices is attenuated relative to variation in hypothetical choices. In specifications adjusting for multiple hypothetical evaluations at once, the standard hypothetical choice consistently enters with the largest positive coefficient, but vicarious choice also receives significant positive weight. We report these estimates in Appendix Table A5 for the multivariate specifications shown in Columns (2)–(3) of Table 3 and in Appendix Table A6 for univariate mappings involving a single hypothetical variable.

4.2.2 Endogenous treatments

We next simulate an application with endogenous pricing: firms learn about demand through hypothetical WTP surveys and set prices to maximize profits. The analyst typically would not have access to those surveys, so we do not include the WTP responses in the vector $\boldsymbol{H}$. Because WTPs exhibit distinctive patterns compared with the other hypothetical responses, the simulations present our method with a serious challenge. We describe the simulated price-setting procedure in Appendix F.6.

Because treatment varies within the simulated dataset, one could in principle employ

[50] For the dichotomized WTP choice, which fails the overlap diagnostic, the procedure yields an accurate estimate when all snacks are observed at the high price, but exhibits a larger bias when all snacks are observed at the low price (Column (6)). This finding is not surprising considering that more extrapolation is required when predicting the high price.

standard real-choice methods. However, a regression that controls for the treatment and physical characteristics exhibits substantial endogeneity bias: the estimate of τ is -0.028 in the median simulation.

Diagnostics that speak to the validity of Assumption 3C, and consequently to whether the two-step procedure relies primarily on ambient variation, are less favorable in these simulations because the variation in the hypothetical evaluations across treatment states is substantial.[51] Nevertheless, for the same Step 1 specifications as in Table 3, we obtain estimates ranging from -0.070 to -0.081, which remains close to ground truth (-0.075), as shown in Table A7. Regressions of the outcome on a treatment indicator, controlling for hypothetical evaluations of both treatment states in addition to physical characteristics of the snacks, yield treatment effect estimates between -0.069 and -0.077 depending on the specification, as shown in Table A8. Consequently, in this application, the two-step procedure may work well because treatment is unconfounded conditional on the hypothetical evaluations—a case subsumed by Proposition 1 rather than Proposition 2. As shown in Appendix F.6, results are similar when accounting for endogeneity with the sample selection correction proposed in Section 2.3.2. If we did not know ground truth, the insensitivity of the estimates to the model of treatment assignment would be reassuring. The bounding method yields ranges of $[-0.158, -0.035]$ and $[-0.163, -0.030]$ for the two low-dimensional specifications of Table 3.[52] While these ranges are wide, they reflect ground truth better than estimates that take hypothetical choices at face value (as in Appendix Table A4(a)).

4.3 Additional benefits

Next, we explore two additional benefits of our method mentioned at the end of Section 2.3.1. To emphasize that these advantages are present even when standard real-choice methods are applicable, we focus primarily on simulated environments with random treatment assignment.

[51] Hypotheticals have decent explanatory power (condition (iii)): in the median simulation, for the specification corresponding to Column (2) in Table A7, a regression of $Y_j(0)$ on $\boldsymbol{H}_j(0)$ and $\boldsymbol{X}_j$ among the untreated observations yields $R^2 = 0.43$. However, the variation in hypotheticals is not predominantly ambient (condition (i)): in the median simulated simulation, for each of the four hypothetical choice variables, treatment accounts for 21–74% of this variation as measured by the ratio $\frac{\text{var}(W_j(H_j(1)-H_j(0)))}{\text{var}(H_j(0))}$.

[52] These bounds impose Invariant Mapping, linearity, monotonicity of demand in price for every snack, and natural limits on potential outcomes ($\underline{Y} = 0$, $\overline{Y} = 1$). With only the natural limits, the bounding method yields a wide range for the average treatment effect, $[-0.543, 0.457]$.

Heterogeneity in treatment effects Even in settings with random treatment assignment, our method may reveal heterogeneity in treatment effects at a level of granularity that is difficult to achieve using real choices alone. To explore this possibility, we estimate the treatment effect for each unit j, $\hat{\tau}_j = \hat{Y}_j(1) - \hat{Y}_j(0)$, by implementing Step 2 of our method separately for each snack based on the same Step 1 specification as in Table 3 Column (3). Through comparisons with benchmarks, we evaluate how well these estimates capture the heterogeneity in actual effects, $\tau_j = Y_j(1) - Y_j(0)$. A full analysis of this issue appears in Appendix F.7, where we consider multiple treatment assignment processes, performance measures, and benchmarks. Here, we focus on results involving random treatment assignment (wherein the price is high for 94 of 189 randomly selected snacks and low for the others), two performance measures (R^2 for a regression of τ_j on $\hat{\tau}_j$ and mean squared error relative to predicting with the ATE), and two natural benchmarks. These benchmarks measure how much variation in treatment effects is explained by observable characteristics other than hypotheticals. The first relates actual treatment effects to linear terms for all covariates other than hypothetical responses (i.e., physical characteristics). The second involves a causal forest (Wager and Athey 2018) that conditions on treatment and characteristics nonlinearly but only employs data on demand for each snack at a single price. The second would be feasible in an actual application; the first would not. To be clear, neither benchmark represents the "ideal" estimator, which would achieve an R^2 of unity and an MSE of zero.

In this application, hypothetical evaluations contain substantially more information about variation in treatment effects than physical characteristics. Specifically, the R^2 is significantly higher for our estimator than for both benchmarks (0.21 versus 0.05 for the first benchmark and 0.00 for the second in the median simulated sample), and the relative mean squared error is lower (0.83 versus 0.95 for the first benchmark and 1.04 for the second in the median simulation). Our method similarly outperforms the benchmarks in simulations with no variation in real choices, and with endogenous variation.

Gains in precision Our approach can yield more precise estimates of treatment effects than common estimators, especially when there are few treated observations (or few untreated observations). With common estimators, performance deteriorates when the fraction treated is far from half because the smaller of the treatment and control groups dominates the variance. With our method, precision does not necessarily deteriorate because Step 1 of our method pools observations for both treatment states, while Step 2 uses

hypothetical evaluations for both treatment states for every setting. Pooling of observations for both treatment states can also reduce the variance of the estimator when the treatment and control groups are of similar sizes. We illustrate this point in Appendix F.8 by varying the proportion of snacks observed at the high price across simulations with random treatment assignment. As expected, the standard deviation of the difference between the mean choice frequencies in the low-price and high-price groups is U-shaped in the fraction of high-price observations. In contrast, the precision of our low-dimensional estimator is largely independent of the proportion treated. Consider, for instance, the estimator using the standard hypothetical choice variable as the only predictor. The standard deviation of the estimation treatment effect is 76% lower with our method than with the difference-in-means for a 50-50 split between low-price and high-price observations, 82% lower for a 75-25 split, and 91% lower for a 90-10 split.

5 Conclusion

We have proposed estimators that infer the causal effects of treatments on choices by combining real choices and hypothetical evaluations based on a new Invariant Mapping assumption. We have explored the implications of this method theoretically and demonstrated that our estimators yield promising results in both a field application and a laboratory application. While the approach is not universally applicable, it augments the causal inference toolbox for suitable applications. Further, if our premise—that cognitive processes reduce all external conditions to the internal motivations that determine choice—is correct, then in principle the relationship may be stable across a broad domain that encompasses many diverse applications. In that case, it may be possible to develop more general mappings between hypothetical responses and real choices, thereby avoiding the need to reestimate those relationships for every application. Because the method generates empirical estimates of the relationship between hypothetical evaluations and real choices each time it is applied, it may be possible, over time, to identify domains within which those relationships are portable.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". In: *Journal of the American Statistical Association* 105.490, pp. 493–505.

Ajzen, Icek, Thomas C. Brown, and Franklin Carvajal (2004). "Explaining the Discrepancy between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation". en. In: *Personality and Social Psychology Bulletin* 30.9, pp. 1108–1121.

Athey, Susan, Raj Chetty, Guido W. Imbens, and Hyunseung Kang (2024). "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely". In: *Review of Economic Studies*.

Athey, Susan, Raj Chetty, and Guido Imbens (2020). "Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes". In: *arXiv:2006.09676*.

Athey, Susan, Guido W. Imbens, and Stefan Wager (2018). "Approximate residual balancing: debiased inference of average treatment effects in high dimensions". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4, pp. 597–623.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones (2012). "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" en. In: *American Economic Review* 102.5, pp. 2083–2110.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot (2014). "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference". en. In: *American Economic Review* 104.9, pp. 2698–2735.

Bernheim, B. Douglas, Kristy Kim, and Dmitry Taubinsky (2024). *Welfare and the Act of Choosing*. en. Tech. rep. w32200. National Bureau of Economic Research.

Berry, Steven, James Levinsohn, and Ariel Pakes (2004). "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market". In: *Journal of Political Economy* 112.1, pp. 68–105.

Blackburn, McKinley, Glenn W. Harrison, and E. Elisabet Rutström (1994). "Statistical Bias Functions and Informative Hypothetical Surveys". en. In: *American Journal of Agricultural Economics* 76.5, pp. 1084–1088.

Blamey, R. K., J. W. Bennett, and M. D. Morrison (1999). "Yea-Saying in Contingent Valuation Surveys". In: *Land Economics* 75.1, pp. 126–141.

Boyle, Kevin J., Sapna Kaul, and Christopher F. Parmeter (2015). "Meta-analysis: Econometric Advances and New Perspectives Toward Data Synthesis and Robustness". en. In: *Benefit Transfer of Environmental and Resource Values*. Ed. by Robert J. Johnston,

John Rolfe, Randall S. Rosenberger, and Roy Brouwer. Vol. 14. Dordrecht: Springer Netherlands, pp. 383–418.

Brandts, Jordi and Gary Charness (2009). “The Strategy versus the Direct-response Method: A Survey of Experimental Comparisons”. en. In: *mimeo*.

Carson, Richard T., Theodore Groves, and John A. List (2014). “Consequentiality: A Theoretical and Experimental Exploration of a Single Binary Choice”. In: *Journal of the Association of Environmental and Resource Economists* 1.1/2, pp. 171–207.

Champ, Patricia A., Richard C. Bishop, Thomas C. Brown, and Daniel W. McCollum (1997). “Using Donation Mechanisms to Value Nonuse Benefits from Public Goods”. en. In: *Journal of Environmental Economics and Management* 33.2, pp. 151–162.

Conley, T. G. (1999). “GMM estimation with cross sectional dependence”. In: *Journal of Econometrics* 92.1, pp. 1–45.

Cummings, Ronald G. and Laura O. Taylor (1999). “Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method”. en. In: *American Economic Review* 89.3, pp. 649–665.

Diamond, Peter A. and Jerry A. Hausman (1994). “Contingent Valuation: Is Some Number Better than No Number?” en. In: *Journal of Economic Perspectives* 8.4, pp. 45–64.

Fang, Zheng, Andres Santos, Azeem M. Shaikh, and Alexander Torgovitsky (2023). “Inference for Large-Scale Linear Systems With Known Coefficients”. en. In: *Econometrica* 91.1, pp. 299–327.

Fox, John A., Jason F. Shogren, Dermot J. Hayes, and James B. Kliebenstein (1998). “CVM-X: Calibrating Contingent Values with Experimental Auction Markets”. en. In: *American Journal of Agricultural Economics* 80.3, pp. 455–465.

Hausman, Jerry (2012). “Contingent Valuation: From Dubious to Hopeless”. en. In: *Journal of Economic Perspectives* 26.4, pp. 43–56.

Heckman, James J (1976). “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models”. In: *Annals of Economic and Social Measurement*. Vol. 5. 4, pp. 457–492.

Heckman, James J. (1979). “Sample Selection Bias as a Specification Error”. In: *Econometrica* 47.1, pp. 153–161.

Huck, Steffen and Imran Rasul (2011). “Matched fundraising: Evidence from a natural field experiment”. en. In: *Journal of Public Economics*. Charitable Giving and Fundraising Special Issue 95.5, pp. 351–362.

Infosino, William J. (1986). “Forecasting New Product Sales from Likelihood of Purchase Ratings”. In: *Marketing Science* 5.4, pp. 372–384.

Jackman, Simon (1999). “Correcting surveys for non-response and measurement error using auxiliary information”. en. In: *Electoral Studies* 18.1, pp. 7–27.

Jacquemet, Nicolas, Robert-Vincent Joule, Stéphane Luchini, and Jason F. Shogren (2013). “Preference elicitation under oath”. en. In: *Journal of Environmental Economics and Management* 65.1, pp. 110–132.

Jamieson, Linda F. and Frank M. Bass (1989). “Adjusting Stated Intention Measures to Predict Trial Purchase of New Products: A Comparison of Models and Methods”. en. In: *Journal of Marketing Research* 26.3, pp. 336–345.

Johansson-Stenman, Olof and Henrik Svedsäter (2012). “Self-image and valuation of moral goods: Stated versus actual willingness to pay”. en. In: *Journal of Economic Behavior & Organization* 84.3, pp. 879–891.

Johnston, Robert J., Mahesh Ramachandran, and George R. Parsons (2015). “Benefit Transfer Combining Revealed and Stated Preference Data”. en. In: *Benefit Transfer of Environmental and Resource Values*. Ed. by Robert J. Johnston, John Rolfe, Randall S. Rosenberger, and Roy Brouwer. Vol. 14. Dordrecht: Springer Netherlands, pp. 163–189.

Kang, Min Jeong, Antonio Rangel, Mickael Camus, and Colin F. Camerer (2011). “Hypothetical and Real Choice Differentially Activate Common Valuation Areas”. en. In: *Journal of Neuroscience* 31.2, pp. 461–468.

Karlan, Dean and John A. List (2007). “Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment”. en. In: *American Economic Review* 97.5, pp. 1774–1793.

Katz, Jonathan N. and Gabriel Katz (2010). “Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout”. In: *American Journal of Political Science* 54.3, pp. 815–835.

Kessler, Judd B. and Alvin E. Roth (2012). “Organ Allocation Policy and the Decision to Donate”. en. In: *American Economic Review* 102.5, pp. 2018–2047.

Krueger, Alan B. and Ilyana Kuziemko (2013). “The demand for health insurance among uninsured Americans: Results of a survey experiment and implications for policy”. en. In: *Journal of Health Economics* 32.5, pp. 780–793.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva (2015). “How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments”. en. In: *American Economic Review* 105.4, pp. 1478–1508.

Lee, Lung-Fei (1983). “Generalized Econometric Models with Selectivity”. In: *Econometrica* 51.2, pp. 507–512.

Levy, Ifat, Stephanie C. Lazzaro, Robb B. Rutledge, and Paul W. Glimcher (2011). “Choice from Non-Choice: Predicting Consumer Preferences from Blood Oxygenation Level-Dependent Signals Obtained during Passive Viewing”. en. In: *Journal of Neuroscience* 31.1, pp. 118–125.

List, John A. (2011). “The Market for Charitable Giving”. en. In: *Journal of Economic Perspectives* 25.2, pp. 157–180.

List, John A. and Craig A. Gallet (2001). “What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?” en. In: *Environmental and Resource Economics* 20.3, pp. 241–254.

List, John A., Azeem M. Shaikh, and Atom Vayalinkal (2023). “Multiple testing with covariate adjustment in experimental economics”. en. In: *Journal of Applied Econometrics* 38.6, pp. 920–939.

List, John A. and Jason F. Shogren (1998). “Calibration of the difference between actual and hypothetical valuations in a field experiment”. en. In: *Journal of Economic Behavior & Organization* 37.2, pp. 193–205.

— (2002). “Calibration of Willingness-to-Accept”. en. In: *Journal of Environmental Economics and Management* 43.2, pp. 219–233.

Little, Joseph and Robert Berrens (2004). “Explaining Disparities between Actual and Hypothetical Stated Values: Further Investigation Using Meta-Analysis”. en. In: *Economics Bulletin* 3.6, pp. 1–13.

Loomis, John, Kerri Traynor, and Thomas Brown (1999). “Trichotomous Choice: A Possible Solution to Dual Response Objectives in Dichotomous Choice Contingent Valuation Questions”. In: *Journal of Agricultural and Resource Economics* 24.2, pp. 572–583.

Louviere, Jordan J., David A. Hensher, and Joffre D. Swait (2000). *Stated Choice Methods: Analysis and Applications*. en. Cambridge University Press.

Lusk, Jayson L. and Ted C. Schroeder (2004). “Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks”. en. In: *American Journal of Agricultural Economics* 86.2, pp. 467–482.

Mansfield, Carol (1998). “A Consistent Method for Calibrating Contingent Value Survey Data”. In: *Southern Economic Journal* 64.3, pp. 665–681.

Manski, Charles F. (1990). “Nonparametric Bounds on Treatment Effects”. In: *The American Economic Review* 80.2, pp. 319–323.

Mas, Alexandre and Amanda Pallais (2017). “Valuing Alternative Work Arrangements”. en. In: *American Economic Review* 107.12, pp. 3722–3759.

— (2019). “Labor Supply and the Value of Non-Work Time: Experimental Estimates from the Field”. en. In: *American Economic Review: Insights* 1.1, pp. 111–126.

Morwitz, Vicki G., Joel H. Steckel, and Alok Gupta (2007). “When do purchase intentions predict sales?” en. In: *International Journal of Forecasting* 23.3, pp. 347–364.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead (2005). “A meta-analysis of hypothetical bias in stated preference valuation”. In: *Environmental and Resource Economics* 30.3, pp. 313–325.

Nelson, Jon P. (2015). “Meta-analysis: Statistical Methods”. en. In: *Benefit Transfer of Environmental and Resource Values*. Ed. by Robert J. Johnston, John Rolfe, Randall S. Rosenberger, and Roy Brouwer. Vol. 14. Dordrecht: Springer Netherlands, pp. 329–356.

Newey, Whitney K. (2009). “Two-step series estimation of sample selection models”. en. In: *Econometrics Journal* 12, S217–S229.

Newey, Whitney K. and Daniel McFadden (1994). “Chapter 36 Large sample estimation and hypothesis testing”. en. In: *Handbook of Econometrics*. Vol. 4. Elsevier, pp. 2111–2245.

Penn, Jerrod M. and Wuyang Hu (2018). “Understanding Hypothetical Bias: An Enhanced Meta-Analysis”. In: *American Journal of Agricultural Economics* 100.4, pp. 1186–1206.

Poe, Gregory L. (2016). “Behavioral Anomalies in Contingent Values and Actual Choices”. en. In: *Agricultural and Resource Economics Review* 45.2, pp. 246–269.

Prentice, Ross L. (1989). “Surrogate endpoints in clinical trials: Definition and operational criteria”. en. In: *Statistics in Medicine* 8.4, pp. 431–440.

Rolfe, John, Roy Brouwer, and Robert J. Johnston (2015). “Meta-analysis: Rationale, Issues and Applications”. en. In: *Benefit Transfer of Environmental and Resource Values*. Ed. by Robert J. Johnston, John Rolfe, Randall S. Rosenberger, and Roy Brouwer. Vol. 14. Dordrecht: Springer Netherlands, pp. 357–381.

Romano, Joseph P. and Michael Wolf (2010). “Balanced control of generalized error rates”. In: *The Annals of Statistics* 38.1, pp. 598–633.

Rothschild, David M. and Justin Wolfers (2011). *Forecasting Elections: Voter Intentions Versus Expectations*. en. SSRN Scholarly Paper ID 1884644. Rochester, NY: Social Science Research Network.

Rothschild, David (2009). “Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases”. In: *Public Opinion Quarterly* 73.5, pp. 895–916.

Smith, Alec, B. Douglas Bernheim, Colin F. Camerer, and Antonio Rangel (2014). “Neural Activity Reveals Preferences without Choices”. en. In: *American Economic Journal: Microeconomics* 6.2, pp. 1–36.

Tusche, Anita, Stefan Bode, and John-Dylan Haynes (2010). “Neural Responses to Unattended Products Predict Later Consumer Choices”. en. In: *Journal of Neuroscience* 30.23, pp. 8024–8031.

Wager, Stefan and Susan Athey (2018). “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. en. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Whitehead, John C., Subhrendu K. Pattanayak, George L. Van Houtven, and Brett R. Gelso (2008). “Combining Revealed and Stated Preference Data to Estimate the Nonmarket Value of Ecological Services: An Assessment of the State of the Science”. en. In: *Journal of Economic Surveys* 22.5, pp. 872–908.

Whitehead, John C., Melissa S. Weddell, and Peter A. Groothuis (2016). “Mitigating Hypothetical Bias in Stated Preference Data: Evidence from Sports Tourism”. en. In: *Economic Inquiry* 54.1, pp. 605–611.

Wiswall, Matthew and Basit Zafar (2018). “Preference for the workplace, investment in human capital, and gender”. In: *The Quarterly Journal of Economics* 133.1, pp. 457–507.

Wuthrich, Kaspar and Ying Zhu (2021). “Omitted variable bias of Lasso-based inference methods: A finite sample analysis”. In: *arXiv:1903.08704*.