CAUSAL INFERENCE FROM HYPOTHETICAL EVALUATIONS

B. Douglas Bernheim[*], Daniel Björkegren[†], Jeffrey Naecker[‡], Michael Pollmann[§]

August 19, 2022

Abstract

This paper develops a method to infer causal effects of treatments on choices, by exploiting relationships between choices and hypothetical evaluations. Under specified conditions, it can recover treatment effects even if the treatment is assigned endogenously and standard estimation methods are poorly suited, or if the treatment does not vary. Additional advantages include more comprehensive recovery of heterogeneous treatment effects and potential improvements in precision. We provide proof of concept by using the approach to estimate the price responsiveness of the demand for snack foods in the laboratory, and the response of contributions to the availability of matching funds on a microfinance website.

KEYWORDS: causal inference, hypotheticals, counterfactuals, machine learning

[*]Stanford University; bernheim@stanford.edu
[†]Brown University; dan@bjorkegren.com
[‡]Google; jnaecker@google.com
[§]Duke University; michael.pollmann@duke.edu

Thank you to multiple 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 especially grateful to Antonio Rangel for his contributions to the earlier project. We are also grateful 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`.

1 Introduction

This paper considers the standard task of inferring the causal effects of a treatment, such as a price or policy intervention, on choices.[1] Applications of this type commonly encounter two types of challenges. First, when treatments are endogenous, correlations between treatment states and choices are potentially spurious. Second, the treatment of interest may be an untested proposal, or it may be rare. For instance, if it is an innovative policy adopted by a single jurisdiction, its effects may be indistinguishable from random variation.

A common approach is to study how observed choices respond to treatment variation arising from arguably exogenous factors (for instance instruments or discontinuities). However, such factors are often difficult to find. Even when they are available, estimates of the causal relationship may be imprecise, particularly if an instrument has a weak connection to the treatment or if there are few observations near a discontinuity. Also, these methods can only identify average treatment effects that are "local" to the units affected by the exogenous factors, which may not coincide with the units of interest. And since these methods exploit observed treatment variation, one cannot use them to evaluate proposed treatments prior to implementation. Other common methods have similar limitations.

One alternative is to ask people, hypothetically, what they would choose under various conditions, an approach commonly called *stated preferences* (for reviews see Shogren, 2005, 2006; Carson and Hanemann, 2005; Carson, 2012). If hypothetical choices were simply noisy measures of real choices, then this approach could solve both challenges, because it does not rely on observed treatments.[2] It would even allow the analyst to recover treatment effects for arbitrary subgroups of the units of observation. Unfortunately, hypothetical choices are systematically biased measures of actual choices (List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005).[3] Still, the fact that these biases are *systematic* suggests that hypothetical choices encode relevant information, and consequently may be good *predictors* of actual choices, even if they are bad predictions. Indeed, the correlation between hypothetical and real choices is usually high.

This paper develops methods for measuring treatment effects by exploiting answers to hypothetical questions. Our approach combines hypothetical responses, for which the treatment is not confounded but outcomes are measured with bias, with observational data on choices, for which the treatment may be confounded but real outcomes are measured

[1] Similar methods are also potentially applicable to settings in which choices pertain to the treatment, and the treatment determines an outcome (conditional on other factors). We briefly outline such applications in Section 2.7. See also Briggs et al. (2020), which complements the current paper by focusing on these alternative settings.

[2] For example, Krueger and Kuziemko (2013) uses hypothetical choices to estimate the price elasticity of demand for health insurance among the uninsured, for whom there is no real choice variation.

[3] The bias typically overstates willingness-to-pay, especially for alternatives that are viewed as more "virtuous."

without bias. We consider subjective responses that aggregate underlying motivations, such as stated preference and hypothetical choices, as well as a variety of responses that capture underlying motivations (such as temptation or social image) that may influence the direction and magnitude of hypothetical bias.

We estimate the predictive relationship between hypothetical responses and real choices in observational data, and then use that relationship to infer the effects of counterfactuals. To be more specific, suppose the treatment of interest, $w \in \{0, 1\}$, varies across settings, indexed by j. Examples include prices varying over a group of related products, or policies varying across jurisdictions. The actual (aggregated) choice outcome for setting j is $Y_j(w)$ in treatment state w. We are interested in the average treatment effect, $\tau = \mathbb{E}(Y_j(1) - Y_j(0))$. However, we observe each setting j only in some realized treatment state $w = W_j$, which may be correlated with potential outcomes. Imagine collecting hypothetical evaluations of the options available in setting j, $\boldsymbol{H}_j(w)$, for both treatment states. First, we estimate a model relating outcomes in the realized treatment states, $Y_j(W_j)$, to the corresponding hypothetical responses, $\boldsymbol{H}_j(W_j)$. For the linear model $Y_j(W_j) = \boldsymbol{H}_j(W_j)\boldsymbol{\beta} + \epsilon_j$, we obtain the OLS estimate $\hat{\boldsymbol{\beta}}$. Second, we use that relationship to predict the outcome for each treatment state. The difference yields an estimate of the treatment effect, $\hat{\tau} = \big(\overline{\boldsymbol{H}(1)} - \overline{\boldsymbol{H}(0)}\big)\hat{\boldsymbol{\beta}}$, which uses the estimated prediction equation to unwind the systematic biases embedded in the average hypothetical responses, $\overline{\boldsymbol{H}(1)}$ and $\overline{\boldsymbol{H}(0)}$. We develop a simple linear estimator suitable for low-dimensional settings, as well as a machine learning estimator suitable for high-dimensional settings. The latter is based on approximate residual balancing (ARB, Athey et al., 2018), an extension of LASSO. We also outline results for doubly robust and nonlinear estimators.[4]

As long as the predictive relationship is stable, this method should yield unbiased estimates of treatment effects. We (i) articulate conditions that would yield stability, and describe the contexts where the approach is applicable, (ii) develop the econometric theory for the estimator, and (iii) provide proof of concept by applying the method to real data involving two separate applications, one in the laboratory, the other in the field. In these applications, the method recovers measures of treatment effects that are close to ground-truth estimates, even under conditions that render standard methods inapplicable.[5]

[4] An accompanying R package is available on Github: `https://github.com/michaelpollmann/hypeRest`.

[5] There are some parallels to studying the relationship between outcomes and hypothetical responses in the literature on stated preference and contingent valuation. A strand on statistical calibration (Kurz, 1974; Shogren, 1993; Blackburn et al., 1994; National Oceanic and Atmospheric Association, 1994; Fox et al., 1998; List and Shogren, 1998, 2002; Mansfield, 1998) typically treats the individual as the unit of observation, whereas our approach treats the decision problem as the unit of observation. A strand on meta-analyses (Carson and Hanemann, 2005; List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005) evaluates the effects of experimental methods on hypothetical bias. Our approach is related to methods that estimate demand for products by modeling demand for product characteristics (Lancaster, 1966). In essence, we treat underlying motivations as product characteristics and elicit them through survey responses. Our method is similarly related

In our first application, we use our method to estimate the demand for various snacks as a function of prices in a laboratory setting. We ask some participants to decide whether to purchase each snack at prices \$0.25 and \$0.75. Other participants evaluate each snack hypothetically along several dimensions at the low price and the high price. We simulate endogenous price variation by restricting the choice data to a single price for each snack, selected in a manner that introduces correlation with demand. We also simulate a data set with no price variation. We compare the resulting estimates of treatment effects against ground truth estimates based on actual purchase decisions for each snack at both prices (which are observable in a laboratory setting).

In our second application, we use our method to assess the effects of matching provisions on lending through a microfinance platform. The observational data tell us the speed at which each borrower profile attracted funding, and whether a third party offered matching funds. We gathered hypothetical data by asking Amazon Mechanical Turk workers to assess these profiles in both the matched and unmatched states. We compare estimates of treatment effects obtained from applications of our methods with ground truth estimates, which we derived from a controlled experiment on the platform.

These applications highlight four potential advantages of our method, when it is applicable.

First, our method can recover average treatment effects in settings with endogeneity even when standard methods are inapplicable. In both applications, the difference between treated and untreated units yields a biased estimate of the treatment effect, because treatment is endogenously assigned. Standard controls do not help, and instruments are not readily available. Hypothetical choices per se are poor predictions of real choices due to hypothetical biases. We also test adjustments intended to "fix" hypothetical bias by changing the protocol, such as asking respondents to take their choices seriously (as in Cummings and Taylor, 1999), asking about intensity (analogously to Champ et al., 1997), or eliciting beliefs about others' choices (to eliminate image concerns and thereby potentially obtain more honest answers, analogously to Rothschild and Wolfers, 2011). In the snack setting, these alternative protocols instead simply introduce additional biases that in most cases do not reduce the baseline hypothetical bias. However, in both settings, our method yields treatment effect estimates close to the ground-truth estimates.

Second, our method can recover treatment effects even when no unit is treated. If the

to demand estimation approaches that augment real choices with additional data such as hypothetical second choices (Berry et al., 2004; Conlon et al., 2021) or measures of relatedness gathered from surveys (Magnolfi et al., 2022). There is also related work in marketing (Juster, 1964; Morrison, 1979; Infosino, 1986; Jamieson and Bass, 1989; Morwitz et al., 2007), political science (Louviere, 1993; Polak and Jones, 1997; Ben-Akiva et al., 1994; Jackman, 1999; Alpizar Rodriguez et al., 2003; Katz and Katz, 2010), and neuroeconomics (Smith et al., 2014). See Appendix B for more discussion.

hypothetical evaluations mostly vary over the same range with and without the treatment, we can, in effect, infer choices in a given setting for the unobserved treatment by examining choices in other settings that evoke similar hypothetical evaluations without the treatment. In our snack application, we find that evaluations vary over a wider range (one that includes less positive responses) with the high price than with the low price. As a result, if we assume only high prices are observed, our method yields estimates of price effects close to ground truth; if we assume only low prices are observed, estimates remain reasonable but are further from the truth.

Third, our method yields more comprehensive measures of heterogeneous treatment effects than standard approaches. In fact, it can recover treatment effects for arbitrary subgroups, and does not require random treatment variation. In our snack application, observable characteristics capture only a small fraction of the underlying heterogeneity in treatment effects. We find that the finer measures of response heterogeneity uncovered by our method can dramatically increase simulated profits in a price setting exercise. Alternately, they may cover groups of particular interest, in contrast to standard methods which measure only the local average treatment effects (LATEs) among compliers. In the microfinance setting, estimates of the treatment effect among compliers (LATE) obtained through our method line up with the ground truth inferred from experimental instrumental variables. However, the experiment cannot identify the effects on other compliance groups, nor the average treatment effect (ATE). Our estimates suggest that matching is twice as effective for the profiles that are not currently matched on the website (compliers) than for those that are already matched (always takers), possibly because the profiles that attract matches also attract loans on their merits. It follows that the platform may be able to raise more funds by modifying the criteria used for match eligibility.

Fourth, we demonstrate that our method can improve the precision of estimated treatment effects even when randomized treatment variation is available, particularly when treatment groups have unbalanced sizes. Because we estimate a single model of the outcome as a function of hypothetical evaluations using all settings, and then use hypothetical data in both treatment states to predict outcomes for every setting, imbalance has no direct impact on the precision of our method. We obtain precise measures of treatment effects even when the treatment is rare (or not observed) in practice.

To be clear, we do not offer this method as a panacea. As we explain, the assumptions that justify our approach are potentially problematic in applications with identifiable features – for example, those for which it is difficult to depict decision problems comprehensively for survey respondents, or to obtain survey samples from populations that sufficiently resemble the decision makers. Nonetheless, in some settings the approach may provide a reliable and cost-effective alternative to field experiments, or it may complement field experiments

by offering a low-cost method for exploring large varieties of treatment possibilities before committing to a particular version.

The paper is organized as follows. The next section describes our approach, provides formal foundations, and discusses the characteristics of appropriate (and inappropriate) applications. Section 3 covers the laboratory application, and Section 4 covers the field application. Section 5 concludes.

2 Method

2.1 The problem

We are interested in the effect of some treatment $w \in \{0, 1\}$ on choices made in a collection of settings j.[6] For each setting j, the outcome $Y_j(w)$ represents an aggregation of people's choices (a sum or average). The objective is to estimate the average treatment effect (ATE):

$$\tau = \mathbb{E}\big(Y_j(1) - Y_j(0)\big)$$

where the expectation is taken over a population of settings.

Each setting has a treatment status $W_j \in \{0, 1\}$, which is selected by someone other than the people who choose outcomes. We are particularly concerned with the case where W_j is endogenous to the potential outcomes $Y_j(w)$, or has no variation (either all observed settings are treated, or all are untreated).

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

Product demand. The analyst seeks to estimate price elasticities for a collection of products (alternatively, for the same product across different markets), accounting for the fact that firms set prices endogenously (Wright, 1928; Schultz, 1938; Stone, 1954). Here, settings correspond to products (alternatively, markets), the treatment is price, and outcomes are purchase decisions by customers. We mimic this setting with laboratory data.[7]

Matching of charitable contributions. The analyst seeks to estimate the effect of matching provisions for contributions to appeals posted on an online platform, accounting for the fact that sponsors choose which appeals to match endogenously. Here, settings correspond to appeals, the treatment is the existence of a match, and the outcomes are donation decisions by the platform's users. For similar applications, see Karlan and List (2007) and Huck and Rasul (2011).

[6] While we focus on environments with binary treatments, our methods are more general.

[7] Our framework applies most directly to settings where choices for different products are made independently, but can accommodate substitution across products with slight modifications. (Specifically, each hypothetical question must specify the price of every good.)

2.2 Our approach

Our approach to causal inference builds on an existing method that uses hypothetical choice data. It corrects for the biases that afflict that method.

The existing method for using hypothetical choices. Imagine that in each setting, we ask people similar to the decision makers of interest what they would choose, hypothetically, under both treatment states. For example, we might ask them if they would hypothetically purchase particular goods at particular prices, or donate to different appeals with or without matches. Using their responses, we could then construct the average hypothetical choice $Y_j^H(w)$ for setting j under treatment w.

The most straightforward way to estimate the ATE for the J settings of interest is to compute the difference in average hypothetical choices between the treatment states:

$$\hat{\tau}_{\text{hyp}} = \overline{Y^H(1)} - \overline{Y^H(0)},$$

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 $w \in \{0, 1\}$ for all settings.

An advantage of this strategy is that it does not require the observed treatments, W_j, to have exogenous variation—or any variation at all. In effect, it makes a counterfactual prediction based on the respondent's mental model of the choice process. Previous studies have used this approach to measure, for example, product demand (see, e.g., Juster, 1964; Morrison, 1979; Infosino, 1986; Jamieson and Bass, 1989), health insurance demand among the uninsured (Krueger and Kuziemko, 2013), and intentions to vote (Jackman (1999) and Katz and Katz (2010)); for reviews, see Shogren (2005, 2006); Carson and Hanemann (2005); Carson (2012).

The main problem with this approach is that hypothetical choices are systematically biased (Cummings et al., 1995; Johannesson et al., 1998; List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005; Blumenschein et al., 2008). For example, people tend to overstate purchases, and they exaggerate their proclivities to take "virtuous" actions, such as donating to charities and purchasing healthy foods.[8]

The proposed approach. Our approach estimates how hypothetical evaluations relate to real choices, and then uses that relationship to undo the biases in hypothetical choices. We consider multiple types of hypothetical evaluations, denoted by vector $\boldsymbol{H}_j(w)$ in setting j,

[8]When surveys are consequential, incentive problems also come into play; see Carson and Groves (2007) and Carson et al. (2011). Biases do not appear to be substantial in all settings, however; see, for example, Abdellaoui et al. (2007) for a within-subject comparison of choices over lotteries and stated (cardinal) preferences over monetary payments.

which may include hypothetical choices $Y_j^H(w)$. The simplest variant of our approach has two steps.

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

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

where 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 characteristics.

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

$$\hat{\tau} = \big(\overline{\boldsymbol{H}(1)} - \overline{\boldsymbol{H}(0)}\big)\hat{\boldsymbol{\beta}}$$

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 $w \in \{0, 1\}$ for all settings.

Because our method uses hypothetical evaluations as *predictors* rather than as *measures* of choices, we are free to use any subjective response that aids prediction. $\boldsymbol{H}_j$ can thus include not only hypothetical choices (which are aggregates of multiple underlying motivations), but also measures of specific motivations, such as the extent 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. In effect, this approach crowdsources beliefs about counterfactual choices, and then calibrates them using real-world data. This calibrated crowdsourcing approach is an alternative to requiring an analyst to model hypothetical choice biases explicitly.

For the sake of concreteness, Table 1 maps this framework into the elements of our two applications (demand for snack foods, and matching provisions for microfinance lending). An accompanying R package for our method is available on Github: `https://github.com/michaelpollmann/hypeRest`.

2.3 Statistical assumptions and properties

Under what conditions does our method yield reasonable estimates? This section lists statistical assumptions that ensure our simple linear estimator for the ATE is consistent and asymptotically normal. In the next section, we explain how each assumption relates to properties of the underlying processes, and outline the characteristics of appropriate applications.

Table 1: Applications

		Demand for Snacks	**Microfinance Matching**
Settings	j	snack items	loan profiles
Treatment	w_j	$\begin{cases} 0 & \text{price \$0.25} \\ 1 & \text{price \$0.75} \end{cases}$	$\begin{cases} 0 & \text{contributions not matched} \\ 1 & \text{contributions matched} \end{cases}$
Outcome	$Y_j(w)$	average purchase frequency for item j given the price associated with w	fundraising velocity for loan profile j within the first 24 hours, given matching condition w
Hypothetical responses	$\boldsymbol{H}_j(w)$	average hypothetical responses to item j given the price associated with w	average hypothetical responses to loan profile j given matching condition w
Conditions			
Overlap Condition		The range of hypothetical responses across all snack items when the price is \$0.25 spans the range when the price is \$0.75	The range of hypothetical responses across all loan profiles when unmatched spans the range when matched
Treatment Assignment Condition		The hypothetical responses $\boldsymbol{H}_j(0)$ and $\boldsymbol{H}_j(1)$ span all information about realized demand that impacts the price of good j	The hypothetical responses $\boldsymbol{H}_j(0)$ and $\boldsymbol{H}_j(1)$ span all information about realized lending that impacts whether loan profile j is matched

Assumption 1. State specific hypothetical evaluations. *Only the hypotheticals evaluated in the same treatment state are relevant for a given potential outcome: for* $w \in \{0, 1\}$,

$$\mathbb{E}\Big(Y_j(w) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}_1, \boldsymbol{H}_j(0) = \boldsymbol{h}_0, \boldsymbol{X}_j = \boldsymbol{x}\Big) = \mathbb{E}\Big(Y_j(w) \mid \boldsymbol{H}_j(w) = \boldsymbol{h}_w, \boldsymbol{X}_j = \boldsymbol{x}\Big)$$

Assumption 2. Invariant mapping. *The mapping between potential outcomes and hypothetical evaluations would be the same in either treatment state:*

$$\mathbb{E}\Big(Y_j(0) \mid \boldsymbol{H}_j(0) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}\Big) = \mathbb{E}\Big(Y_j(1) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}\Big)$$

Assumption 3. Linearity. *The conditional expectations of potential outcomes are linear in the predictors: for* $w \in \{0, 1\}$,

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

These three assumptions, which we have stated in order of increasing restrictiveness, justify the simplest variant of our method in settings where assignment to treatment is random, or has no variation (e.g., one treatment state is an untested proposal).[9] To justify our method in settings with endogenous treatment assignment, we require an additional assumption, which is satisfied automatically when there is no variation in treatment or when it is randomly assigned:

Assumption 4. Unconfoundedness. *Treatment assignment is unconfounded conditional on hypothetical evaluations*:

$$W_j \perp\!\!\!\perp Y_j(0) \mid \boldsymbol{H}_j(0), \boldsymbol{X}_j$$
$$W_j \perp\!\!\!\perp Y_j(1) \mid \boldsymbol{H}_j(1), \boldsymbol{X}_j$$

[9]The unrestricted linear form is given by $\mathbb{E}\Big(Y_j(w) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}_1, \boldsymbol{H}_j(0) = \boldsymbol{h}_0, \boldsymbol{X}_j = \boldsymbol{x}\Big) = \boldsymbol{h}_1\boldsymbol{\beta}_{w,1} + \boldsymbol{h}_0\boldsymbol{\beta}_{w,0} + \boldsymbol{x}\boldsymbol{\gamma}_w$. Assumption 1 implies that $\boldsymbol{\beta}_{0,1} = \boldsymbol{\beta}_{1,0} = 0$ and Assumption 2 implies that $\boldsymbol{\beta}_{0,0} = \boldsymbol{\beta}_{1,1} \equiv \boldsymbol{\beta}$ and $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_1 \equiv \boldsymbol{\gamma}$.

Under the stated assumptions, our linear estimator has the following asymptotic distribution.

Theorem 1. *The parametric estimator $\hat{\tau}$ is consistent for the average treatment effect τ and asymptotically normal:*[10]

$$\sqrt{J}\Big(\hat{\tau} - \tau\Big) \to \mathcal{N}\Big(0, V_\tau\Big)$$

when the data $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j)_{j=1}^J$ are a random sample of independent observations, under Assumptions 1, 2, 3, and 4, as well as the standard regularity conditions.

Linearity of conditional expectations (Assumption 3) allows us to extrapolate the relationship between Y and $\boldsymbol{H}$ to unobserved values of $\boldsymbol{H}$. Alternately, we can replace linearity with an overlap condition (Assumption 5, see Appendix D.2). As extensions, we show how this modification can enable nonparametric and machine learning estimators, which may perform better with high dimensional hypothetical evaluations, or when hypothetical biases are nuanced.

2.4 Characteristics of suitable applications

In this section, we explore the plausibility of Assumptions 1, 2, and 4, and identify the types of applications that might satisfy them.

Any decision problem involves choosing from a menu of options. When someone makes a choice, their brain maps each option to a bundle of "motivational attributes" (e.g., the degree to which the option addresses hunger, social approval, and so forth). We can therefore think of the individual as choosing from a "psychological menu" containing bundles of motivational attributes. The central premise of our approach is that if two decision problems map to the same psychological menu, the options a person would select in each problem map to the same item on that menu (or to one that is equally preferred). In that sense, external conditions influence choices only to the extent they change internal psychological motivations.

[10] The formula for the variance matrix is:

$$\begin{aligned} V_\tau = {} & \mathbb{E}\Big((\tau - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta})^2\Big) \\ & + \mathbb{E}\Big(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)\Big)\boldsymbol{V}^{\text{ols}}\mathbb{E}\Big(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)\Big)^T \\ & - 2\mathbb{E}\Big(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)\Big)\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}_j^T(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta})(\tau - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta})\Big), \end{aligned}$$

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 $\boldsymbol{\delta} = [\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T]^T$. $\boldsymbol{V}^{\text{ols}} = \mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j(y - \boldsymbol{Z}_j\boldsymbol{\delta})^2\Big)\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}$ 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 D.1 for details.

For the sake of precision, suppose we are concerned with the choice of $y \in \mathcal{Y}$ (such as the amount purchased of a given item) in a variety of settings j (such as items within a category) and treatment states w (such as price). Each such decision problem induces a menu of motivational attribute bundles, $\{\boldsymbol{\theta}_j(y,w)\}_{y\in\mathcal{Y}}$ (where $\boldsymbol{\theta}_j(y,w)$ is the bundle for option y). If there are two settings j and j' along with treatments w_j and $w_{j'}$ for which $\boldsymbol{\theta}_j(y,w_j) = \boldsymbol{\theta}_{j'}(y,w_{j'})$ for all $y \in \mathcal{Y}$, then our premise is that a person would choose the same value of y in either.[11]

If we replaced $\boldsymbol{H}_j(w)$ with variables $\Theta_j(w)$ that govern the relationship between y and $\boldsymbol{\theta}_j(y,w)$,[12] Assumptions 1 and 2 would simply restate our central premise that decisions depend only on internal psychological motivations for the problem at hand.[13] Assumption 4 would also follow, because the psychological menu encompasses all of the decision-relevant information the treatment selector might consider.

We think of hypothetical evaluations $\boldsymbol{H}$ as proxies for Θ. Our approach requires those proxies to be "adequate," so that the predictive relationship between y and $\boldsymbol{H}$ (potentially conditioning on x) remains stable.In the rest of this section, we elaborate on the characteristics of applications to which our method applies, and in the process clarify the requirement that $\boldsymbol{H}$ is an "adequate" proxy for Θ.

Conditional on the treatment, each outcome is an aggregate of unitary human choices Our methods rely on respondents to evaluate factors that predict outcomes. In principle, one could elicit subjective predictors for any type of outcome and deploy our method. However, if the outcome results from technological or biological processes, respondents may not be sufficiently aware of how the outcome is determined. As an example, suppose the objective is to measure the effect of water purification on health status. Imagine asking community members to predict community health status as a function of treatment status and local conditions in order to apply our method. Community members' predictions may correlate poorly with actual outcomes. In particular, if people are better at predicting outcomes in one of the counterfactual treatment states (e.g., the baseline without treatment), then the predictive relationship would differ between the two states because the noisiness of the predictor differs, As a result, Assumption 2 will fail. Additionally, if treatment is assigned in part based on expert advice (e.g., from public health professionals), it may be based on far better information about the treatment-contingent outcome than survey respondents'

[11] When there are only two options on the menu, $\mathcal{Y} = \{0,1\}$, for instance "buy" ($y = 1$) and "don't buy" ($y = 0$), the relevant information can alternatively be described by the difference in motivational states $\boldsymbol{\theta}_j(1,w) - \boldsymbol{\theta}_j(0,w)$. This simplification is akin to "normalizing the outside option."

[12] For example, if y is continuous and each motivational attribute is linear in y, the first derivatives of $\boldsymbol{\theta}_j(y,w)$ would suffice. There is a close analogy to using price and income as sufficient statistics for all available bundles in standard demand curve estimation.

[13] Technically, stochastic variation would be de minimis, and conditioning on x would be unnecessary.

predictions, in which case Assumption 4 will also fail.

Similar issues arise when the outcome results from a collection of interacting decisions rather than unitary choices. For example, if the previous example had concerned the effect of the minimum wage on equilibrium employment, rather than water purification, the same considerations would come into play. However, our method might be suitable for analyzing partial equilibrium effects of a minimum wage on job search by prospective employees, and, separately, on managers' hiring practices.

Hypothetical evaluations adequately proxy for motivations. The adequacy with which hypothetical evaluations proxy for motivations depends on a number of considerations, some of which the analyst controls, at least to a degree.

Similarity of decision makers and evaluators. Evaluators who more closely resemble the decision makers are likely to have better information about their motivations. That consideration argues for sampling respondents from the population that makes choices, with minimal temporal separation. One caveat is that the relationship between real choices and hypothetical evaluations could be distorted if respondents' real choices under the treatment W_j influenced their hypothetical evaluations (for example, through anchoring or ex post rationalization). This confound is less of a concern when the decisions of interest are differentiated, or are less memorable. For example, no two borrower profiles on the microlending platform we study are alike. It is unlikely that our respondents had previously made decisions about those particular profiles, or remembered them if they had. 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 (for example, just-hired employees who have not yet made 401(k) elections).

Naturalism and familiarity. Hypothetical evaluations are more likely to be informative when descriptions of the choice scenarios bring all the relevant information to mind. In some applications, these scenarios may be so standard that a short hypothetical description suffices (as in our first application, which involves purchases of common snack foods). In others, it may be possible to depict the choice scenarios naturalistically (as in our second application, which involves online microfinance lending). Our method is less likely to work when the study examines choices that are unfamiliar or too complex to fully represent when gathering hypothetical evaluations. For example, hypothetical automobile purchase decisions cannot include test drives.

Spanning of motivational attributes. We require that the set of hypothetical evaluations is sufficiently rich to span the factors underlying the available motivational attribute bundles. Spanning can be achieved with a collection of hypothetical evaluations that reflect composites of motivational attributes, so our method does not require a definitive catalog of motivational

attributes, nor does it rely on pairing each individual attribute with a matching proxy. Using a rich set of hypothetical evaluations, including broad composite evaluations (such as hypothetical choices) along with a variety narrowly focused evaluations (such as the intensity of temptation an option evokes), helps to ensure that the hypothetical evaluations span the information that Θ contains.[14]

To understand the logic of the spanning requirement, imagine that the hypothetical evaluations H are each linear functions of Θ. In that case, any linear function of H is implicitly a linear function of Θ. Now imagine that y is also a linear function of Θ. The purpose of the spanning condition is to ensure that, by appropriately reweighting the elements of H (as a regression would do), we can reproduce any linear function of Θ, including the one that describes y. In that case, one loses neither information nor functional flexibility when replacing Θ with H. This logic has parallels in the literature on linear factor models, which we discuss in Section 2.7.

Even when hypothetical evaluations span the underlying space of motivational attributes, the empirical relationship between Θ and $\boldsymbol{H}$, and hence the relationship between y and $\boldsymbol{H}$, may be unstable (contrary to Assumption 2). One potential reason for instability is that the reporting biases affecting $\boldsymbol{H}$ may vary across settings. We can address this source of instability by expanding $\boldsymbol{H}$ so that it also spans the motivations that impact reporting biases, such as the extent to which others would approve of each response.[15] Second, the relationship between $\boldsymbol{H}$ and Θ may depend on extraneous factors, such as measurement error. We discuss the particular case of sampling error, including standard instrument variables solutions, in Section 2.7.

Spanning of outcome-relevant information used to select treatment. In the case where treatment is selected endogenously, we also need a second form of spanning to satisfy Assumption 4 (unconfoundedness): hypothetical evaluations span the outcome-relevant information used to select the treatment. This condition is more plausible when treatment selection is based on limited information about outcomes, such as general attitudes rather than precise econometric forecasts of behavior. For example, a retailer may set prices based on consumer surveys that are analogous to hypothetical evaluations, rather than causal estimates.

[14] One does not actually need $\boldsymbol{H}$ to subsume *all* the information contained in Θ: a natural possibility is that people answer hypothetical questions by envisioning *typical* decision conditions, rather than the *specific* conditions that give rise to the observed value of y and the associated latent value of Θ. 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 or 2. See Appendix C for an elaboration of this possibility.

[15] Dependence of reporting biases on the motivational attributes of the options themselves (e.g., a tendency to exaggerate the inclination to take a socially approved action) does not necessarily overturn the ability to reexpress any function of Θ as a function of $\boldsymbol{H}$, although it could (for example, in the one-dimensional case, if the relationship between H and Θ becomes non-monotonic).

The condition also may also hold in applications where the attitudes that drive outcomes evolve more rapidly than treatments. Because the selection of treatment W_j necessarily occurs before the choice of $Y_j(W_j)$, the treatment selector's information concerning $Y_j(w)$ is not comprehensive. Any endogeneity arising from the selection of W_j would have to result from persistent influences on choice, which hypothetical evaluations more readily capture. Thus, confounds are weaker where attitudes evolve more rapidly (so that the informativeness of outcome-relevant information decays more quickly) and treatments are hard to change (so they are less tailored to current attitudes).[16] As an example, one may be interested in estimating the effects of policies selected by state legislatures. Because legislators' decisions depend on the behavior and preferences of their constituents, such policies are generally endogenous. However, a legislator may focus on limited types of information, such as their constituents' policy preferences as measured by public opinion polls, rather than forecasts of their policy-contingent choices. Furthermore, because legislative processes exhibit inertia, the legislators' information at the time of adoption may become increasingly "stale" as the public's attitudes evolve. If legislators relied only on old public opinion polls, then current polls may well subsume all the information about current outcomes that was relevant for legislators' policy decision.

2.5 Diagnostic Checks and Robustness

All nonexperimental methods of causal inference rely on untestable assumptions, and ours is no exception.[17] As with standard methods, one can nevertheless use diagnostics to obtain evidence on the method's credibility in a given application (Athey and Imbens, 2017). We outline some diagnostics here, and then deploy them in our applications.

Overlap. Extrapolations of the relationship between y and $\boldsymbol{h}$ to values $\boldsymbol{h} = \boldsymbol{H}_j(1 - W_j)$ outside the observed range of variation for $\boldsymbol{H}_j(W_j)$ lean heavily on functional form assumptions (specifically, Assumption 3). It follows that our method is more applicable when the ranges of variation for the hypothetical evaluations $\boldsymbol{H}_j(W_j)$ and $\boldsymbol{H}_j(1 - W_j)$ largely overlap – in other words, when the effect of the treatment on motivations is not too large relative to other sources of variation in motivations. One can assess whether motivational states in the realized treatment states span those in the unrealized states by examining the overlap between the marginal distributions of $\boldsymbol{H}_j(W_j)$ and $\boldsymbol{H}_j(1 - W_j)$. When the overlap is high, the analysis is less sensitive to assumptions about functional form, and one can

[16] See Appendix C for a more fully detailed illustration of our interpretation.

[17] For example, difference-in-differences relies on parallel trends in the post-treatment period; regression discontinuity designs rely on continuity of the potential outcomes at the threshold; instrumental variables estimation relies on the exclusion restriction and absence of defiers; structural modeling relies on assumptions about the invariance of the structure of the model under counterfactuals; etc.

dispense with functional form assumptions entirely; see Appendix D.2, where we replace Assumption 3 with Assumption 5.

Stability across specifications. As with standard methods, results that are robust across progressively richer specifications may instill greater confidence. For example, suppose the analyst arrives at five highly relevant motivations for the decisions of interest, plus five others of possible importance, and elicits hypothetical evaluations for each. If the results are similar when using the five highly relevant motivations versus the full set of ten, then it is plausible that adding proxies for additional motivations (which the analyst deems even less relevant) beyond the set of ten would also have little if any effect. In that case, one can have greater confidence that the assumptions hold for the elicited set of motivations.

Selecting skilled evaluators. Another diagnostic strategy is to rely on the subset of survey respondents whose hypothetical evaluations predict real decisions most accurately. Focusing on a single element of $\boldsymbol{H}$, we define the *latent response quality*, r_{kj}, for respondent k's evaluation of setting j as the correlation between k's evaluations and outcomes *for other settings* $j' \neq j$. For the purpose of estimating treatment effects, one can set a quality threshold r^* and drop all observations with latent quality below this threshold, $r_{kj} \leq r^*$.[18] Because this strategy reduces the number of evaluations per setting, it may be appropriate to remove any attenuation bias by replacing Step 1 of our method with instrumental variables on split samples (for instance, Fuller, 1987).[19] By varying r^*, the analyst can check whether poor quality evaluations drive the results. We deploy this strategy in Section 4.4.

2.6 Potential Advantages

When our methods are applicable, they offer several potential advantages.

Estimation of treatment effects without quasiexperimental variation. Standard methods of causal inference require variation in treatment that is either related to an identifiable exogenous variable, unrelated to trends, or discontinuous. Our methods require none of these assumptions. Indeed, because they do not require *any* variation in the observed treatment, they allow causal effects to be estimated for novel treatments that are nothing more than proposals. Even though such treatments are unobserved, the hypothetical evaluations

[18] This leave-one-out correlation r_{kj} avoids overfitting by omitting any direct information concerning the predictive accuracy of k's evaluation for the j-th setting.

[19] In this case we observe a small random sample of $\boldsymbol{H}_{kj}(w)$ rather than (approximately) the population measure $\boldsymbol{H}_j(w)$. Under our assumptions, $\mathbb{E}(\boldsymbol{H}_{kj}(w)|\boldsymbol{H}_j(w)) = \boldsymbol{H}_j(w)$, so this consideration implies that we measure $\boldsymbol{H}_j(w)$ with classical measurement error. We randomly split responses into two equal groups, using one half as an instrument for the other. We obtain a second estimate by reversing the roles of the two subgroups, and then average the two estimates.

they induce may fall within the range of variation for observed treatment states. Under our assumptions, observing these hypothetical evaluations can substitute for observing the treatment implemented.

More flexible estimation of heterogeneous treatment effects. Standard observational methods identify only Local Average Treatment Effects (LATEs), which are specific to the units for which the treatment changes in response to variation in an instrument, or at a discontinuity (the compliers; Imbens and Angrist, 1994). However, treatment effects commonly vary across units (here, across settings). Using our method, one can estimate the ATE for any subgroup of settings $\mathcal{S}$ (defined according to values of our conditioning variables $\boldsymbol{H}_j(1), \boldsymbol{H}_j(0)$, and $\boldsymbol{X}_j$) by calculating $\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\left[\hat{Y}_j(1) - \hat{Y}_j(0)\right]$. Such calculations are possible because one observes both treated and untreated hypothetical responses in all settings.

Gains in precision. Because our approach computes treatment effects using data on all settings in both treatment states, it may yield more precise estimates than conventional methods. Gains are more likely when either the mixture of realized treatment states is lopsided, or instruments exploit only a small component of treatment variation.

Encoding and evaluating nuanced features. Standard causal inference methods are typically constrained to use coarse definitions of treatments and settings. For example, the literature on organ donation focuses on a single feature: whether people are invited to opt in or opt out (Kessler and Roth, 2012, 2014). These studies abstract from many other differences among treatments such as wording and placement of the question (which typically appears somewhere on a driver's license application). Likewise, in our second application, users of a microfinance platform view profiles of people seeking loans, some of whom are eligible for matching funds. While the profiles themselves are short, each includes a photo and a narrative description of the loan's purpose. This information is complex and difficult to encode into a manageable set of variables.

Our approach utilizes respondents to encode complex information such as photos or text into quantitative measures, just as decision makers would encode it into their choices. In addition, this method can isolate the effects of any particular feature as long as one assesses hypothetical responses to treatments that differ only with respect to that feature. Moreover, if treatment selection is related to the nuanced features of each decision task, conditioning on such hypothetical responses may encode the decision-relevant portion sufficiently well to render the treatment unconfounded, even when it would be infeasible to condition on a high dimensional description of the task.

In principle, one can also assess the effects of nuanced features in controlled experiments. In practice, the typical field experiment focuses on a single combination of features; see, for example, the description of experiments conducted by nudge units in (DellaVigna and Linos, 2022). Our methods allow the analyst to explore the treatment space at far lower cost by gathering hypothetical responses to a variety of treatment designs. Experiments can then focus on the most promising designs.

2.7 Extensions and additional details

We have described a simple linear estimator to build intuition. Here we mention more complex variants, which we deploy in our two applications, and which may prove useful more generally. We also relate our methods to existing procedures.

Extensions involving machine learning. One may wish to include a sizable collection of hypothetical evaluations, along with interactions, in $\boldsymbol{H}_j$. Some applications may invoke many types of plausibly interacting motivations. For example, the reported pleasure derived from an option may depend on perceptions of social approval. Hypothetical evaluations may also employ arbitrary scales, so one may also want to include transformations such as quadratic terms. With many predictors, linear estimators may overfit, and machine learning estimators may perform better. Appendix D.2 describes an approach similar to LASSO for cases involving linearity and sparsity in high-dimensional hypothetical evaluations, a variant of approximate residual balancing (ARB, Athey et al., 2018). In Appendix D.4, we provide a doubly robust moment condition for estimation using arbitrary machine learning methods.

Extensions involving general nonparametric estimation. Our approach is not generally tied to parametric assumptions such as linearity, because treatment effects are identified non-parametrically, provided the distributions of $\boldsymbol{H}_j(W_j)$ and $\boldsymbol{H}_j(1-W_j)$ overlap. We develop these observations in Appendix D.3.

Extensions that address sampling error. Our theoretical results treat the hypothetical evaluations contained in $\boldsymbol{H}_j(0)$ and $\boldsymbol{H}_j(1)$ as population statistics, rather than as aggregates based on finite samples. This treatment is appropriate when the number of hypothetical responses is large relative to the number of settings. Given a fixed collection of settings, one can always increase the size of the survey sample. As an alternative, one can think of our analysis as conditional on the finite sample of individuals from whom we elicit hypothetical evaluations. A potential concern then arises because the logic of the unconfoundedness

assumption references average motivational responses for the pertinent population.[20] To address this concern, we can treat the discrepancy as a source of classical measurement error in the covariates, and apply standard corrections. We describe one such correction in Section 2.5.

Relation to linear factor models. Our discussion of motivational attributes suggests that a (linear) factor model could provide a plausible microfoundation for our approach. Suppose setting j under treatment state w induces the menu of motivational attribute bundles $\Theta_j(w)$. The outcome and hypothetical evaluations are linear functions of these latent "factors":

$$\begin{aligned} Y_j(w) &= \boldsymbol{\Theta}_j(w)\boldsymbol{\phi}_Y + \epsilon_{Y,j}(w) \\ H_j^q(w) &= \boldsymbol{\Theta}_j(w)\boldsymbol{\phi}_{H^q} + \epsilon_{H^q,j}(w) \qquad \text{for } r = 1, \ldots, R \end{aligned}$$

where $H_j^q(w)$ is the aggregate response to the q^{th} hypothetical question;

$$\boldsymbol{H}_j(w) = [H_j^1(w), \ldots, H_j^{Q_H}(w)] \in \mathbb{R}^{Q_H}.$$

(For simplicity, we consider a model without fixed characteristics $\boldsymbol{X}_j$; including them is straightforward.) Only Y and $\boldsymbol{H}$ are observed; latent variables include $\Theta_j(w)$ (a row vector) and weights $\boldsymbol{\phi}_Y$ and $\boldsymbol{\phi}_{H^q}$ (column vectors). Linear factor models can provide justification for the regression of Y_j on $\boldsymbol{H}_j(W_j)$ in Step 1 of our method.[21] Our Assumptions 1–4 are satisfied under appropriate restrictions on the error terms ϵ.[22] However, the econometric theory in this paper is agnostic about the preferred microfoundation.

Relation to statistical surrogates. We assume that treatments affect the outcomes of interest only through psychological motivations. Consequently, we treat hypothetical evaluations much like statistical surrogates (for instance, Prentice, 1989; Begg and Leung, 2000; Frangakis and Rubin, 2002; Athey et al., 2020). However, statistical surrogates are observed only for the realized treatment state, whereas we observe hypothetical evaluations for all treatment states. This distinction leads to different assumptions, estimators, and properties.

[20] For example, if the treatment selector knew population statistics to a high degree of precision, measurements based on finite samples may not encompass that information adequately due to sampling variation.

[21] This procedure is similar to using them to justify "weight estimation" of synthetic control methods (Abadie et al., 2010).

[22] State specificity most closely relates to a restriction on the relationship between $\boldsymbol{H}(1-w)$ and $\epsilon_Y(w)$. Invariance of the mapping most closely relates to the invariance of the factor loadings $\boldsymbol{\phi}_Y$ and $\boldsymbol{\phi}_{\boldsymbol{H}}$ w.r.t. treatment w imposed in the notation above. Linearity most closely relates to the linearity of the factor model. Unconfoundedness most closely relates to a restriction on the relationship between W and ϵ_Y.

Treatment as choice. Endogeneity commonly arises because the choice of treatment is correlated with potential outcomes. We focus on applications where the outcomes of interest are choices, and elicit hypothetical evaluations from people resembling those making the *choices that constitute the outcome*. One could alternately model the *choice of the treatment,* by eliciting hypothetical evaluations from people resembling the treatment selectors ("treatment as choice", for example Briggs et al. (2020)).[23] While hypothetical evaluations may be of use for both approaches, the "treatment as choice" approach is potentially suitable for different types of applications. For example, it is applicable when the outcome of interest is not a choice per se, or is the result of a complex process, such as health. However, in many cases it is difficult to survey people resembling those selecting treatments, as they may be specialists and few in number (such as retail price strategists, or charitable matching sponsors). If one tries to survey those who made real treatment choices, evaluations may be subject to anchoring or ex post rationalization. More work is needed to identify the characteristics of applications for which "treatment as choice" approaches yield credible estimates.

3 Application: Snack Demand

To test this approach, we use it to estimate price sensitivities for a collection of goods in a laboratory setting. Study participants make simple purchase decisions for a large collection of familiar snack foods. The 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 ATE 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 datasets from a larger one containing two real observations for each good (demand at both prices), which we use to measure true price responses (ground truth). The structure of our study renders the demand for each item independent of the prices for other items, but the method can accommodate substitution across products with slight modifications.

[23] Under this proposal, hypothetical evaluations may proxy for expectations about outcomes, for instance motivated by a Roy model of selection into treatment. In the context of double robust estimation of treatment effects (Robins and Rotnitzky, 1995; Chernozhukov et al., 2018), our "outcome as choice" approach resembles outcome modeling. In contrast, "treatment as choice" approaches resemble propensity score modeling.

3.1 Procedures and data

Each of 365 subjects was assigned to one of several groups, described below.[24] 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). 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 a complete catalog of the groups along with sample sizes and a screenshot for a representative question, see Appendix E.1 and Figure A1.

3.1.1 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. In observational data we might observe such demand at a single price, possibly 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.

Although the chance of implementing any given choice was low, differences between real and hypothetical purchase frequencies were substantial, and in the expected direction.[25] Moreover, 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.[26] It is not surprising that participants in the "real choice" group viewed their choices as real: they had

[24] We conducted the experiment at the Stanford Economic Research Laboratory (SERL) between November 15, 2010, and October 2, 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.

[25] Consistent with prior findings concerning hypothetical choice bias, average purchase frequencies are significantly higher for hypothetical choices than for real choices. Additionally, the cross-choice-task variance of the purchase frequency is considerably higher for hypothetical choices than for these real choices, and the average price sensitivity implied by the purchase frequencies is much larger for hypothetical choices than for these real choices.

[26] See Appendix E.3 for details.

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.[27]

3.1.2 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.

Several groups made hypothetical choices. The literature on stated preferences explores a variety of protocols for eliciting such choices. 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,[28] 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 of the 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.

3.1.3 Patterns of real choices and implied treatment effects (ground truth)

On average, 28% of people elect to purchase a snack when the price is \$0.25. When the price rises to \$0.75, the purchase frequency declines by 7.5 percentage points ($\tau = -0.075$, standard error 0.004).

Across all item-price combinations, the purchase frequency varies from 0% to 60%. There is also substantial variation in demand conditional on price: the sample standard deviation is 11% with a price of \$0.25 and 9% with a price of \$0.75. Demand for some items is much

[27] Notably, similar conclusions were reached by Carson et al. (2011) 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.

[28] We would like to 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 developed in Cummings and Taylor (1999).

more price-sensitive than for others: the standard deviation of the percentage point change is 6 percentage points. An increase in price from \$0.25 to \$0.75 reduces demand for 85% of our items, increases it for 3% of items, and has no effect for the remaining 12% of items.[29]

3.1.4 Patterns of hypothetical choices

As expected, hypothetical choices exhibit substantial hypothetical bias: the average standard-protocol hypothetical demand across all item-price pairs (31%) overstates real demand (24%) by nearly 7 percentage points (equivalently, by 28%), and we reject the absence of bias ($p \leq 0.001$). Moreover, hypothetical demand exceeds the real demand for 70% of item-price pairs.

The variance of hypothetical demand across all item-price pairs is more than twice that of real demand, a phenomenon we call *hypothetical noise*.[30] As we show in Appendix E.4, hypothetical noise is attributable in significant part to greater systematic variability of population hypothetical demand than of population real demand across choice problems, rather than to differences in sampling variation (which might arise if subjects take hypothetical choices less seriously). A possible explanation is that, when answering hypothetical questions, people naturally exaggerate the sensitivity of their choices to characteristics and conditions.

Together, hypothetical bias and hypothetical noise render standard-protocol hypothetical choices poor predictions of real choices. Even so, there is a strong correlation across items between hypothetical and real purchase frequencies (ρ =0.75), which suggests that hypothetical demand may be a useful predictor of real demand, even if it is not a good prediction. Figure 1 shows this relationship more clearly, using orange squares for item-price pairs with prices of \$0.25, and purple dots for pairs with prices of \$0.75. The relationship between hypothetical and real demand is systematic, and, helpfully for our purposes, stable between treatments.[31]

3.2 Estimation under endogenous treatment assignment

In this section, we mimic observational datasets in which each product is offered at a single, endogenous price (the treatment), and we observe the quantity sold (the outcome). Prices

[29] Given the size of the "real choice" group (30 subjects), some of this variation may reflect sampling uncertainty. Differencing may either amplify or reduce that error, depending on the magnitude of the correlation between choices by the same subject involving the same item but different prices. However, in light of our ultimate success in generating predictions of price sensitivities that are reasonably well-calibrated (see Section 3.4), it is safe to conclude that some significant fraction of the variation in the measured responsiveness to price reflects population variation rather than sample variation.

[30] Similarly, Carson et al. (2011) found that the variance of valuations rises when choices become less consequential.

[31] Visually, lowering the price (from purple dots to orange squares) appears to shift the cloud to the northeast (higher hypothetical and real purchase frequencies) without disturbing the relationship between the variables.

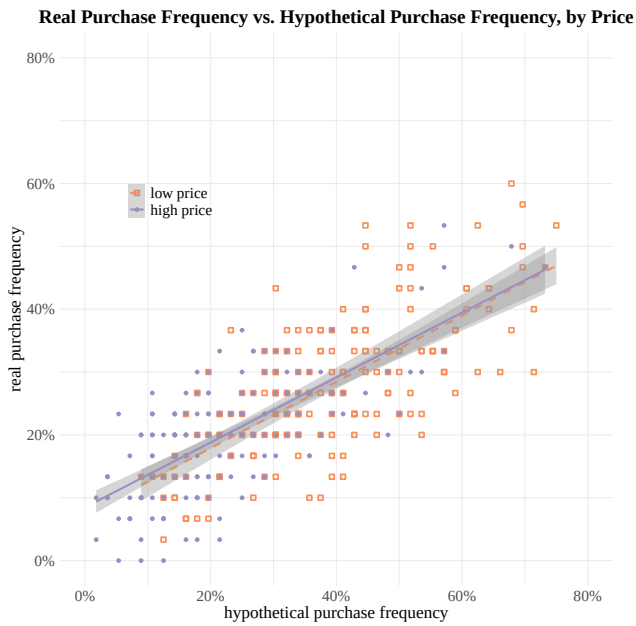


Figure 1: Real vs. Hypothetical Choices

Item-price pairs plotted. Separate regression lines for the $0.25 choices and the $0.75 choices are shown with error bands. A χ^2 test cannot reject the hypothesis that the lines are the same for observations involving items sold at a price of $0.25, and for those involving items sold at a price of $0.75 ($p = 0.58$ assuming independent observations). In the Online Appendix, we show that the curves are approximately linear and similarly overlap when using nonparametric regression.

vary across products, rather than for each individual product.

3.2.1 Endogenous treatment

We select a virtual price for each product that is correlated with potential outcomes, based on respondents' hypothetical willingness to pay (WTP) for it. Specifically, we set

$$W_{jr} = 1\left\{\text{WTP}_j > \epsilon_{jr}\right\},$$

for item j in simulation r. The random shocks ϵ_{jr} are independent draws from a t-distribution with 3 degrees of freedom, with mean set to the median of WTP, and the standard deviation set to that of the WTP distribution.[32] We drop the observation of the real choices at the other price.

Because WTP is strongly correlated with potential outcomes, this procedure generates endogeneity in virtual prices.[33] It simulates an environment in which sellers use consumer

[32]We choose a fat-tailed distribution so that even snacks with extreme WTPs still have a reasonable (if small) chance of being observed at either price.

[33]Appendix Figure A2 plots each snack's actual purchase frequency at the low and high prices (potential outcomes) against the simulated probability it is observed at the high price. There is a strong positive relationship. Alternative assignment mechanisms, such as mimicking the decisions of profit maximizing producers who are exposed to exogenous marginal cost shocks, or assignment based on measured price elasticities, yield qualitatively

surveys to assess the attractiveness of their products when setting prices. Because the analyst typically would not have access to those surveys, we do not include hypothetical WTPs in the vector $\boldsymbol{H}_j$ when deploying our method, except where noted.

3.2.2 Results

We first compare univariate versions of the estimators proposed in this paper to some simple standard estimators discussed in the literature. Table 2 shows median estimates and standard errors for each estimator across simulated samples r. The table also includes the ground truth estimate (Column (1)), i.e., that increasing the price from \$0.25 to \$0.75 changes the proportion of subjects buying the average snack by -0.075 percentage points.

The difference in means (mean of treated minus mean of control, Column (2)) yields an estimated effect of -0.025. As in real applications, endogeneity leads this simple estimator to misstate price sensitivity.

Treating standard hypothetical choices as predictions (i.e., estimating the effect as the mean difference in hypothetical choices, Column (3)) yields an estimated effect of -0.159, which implies a significant bias in the opposite direction. Because hypothetical choices are observed in both treatment states, here the discrepancy arises from hypothetical choice bias rather than from endogenous treatment assignment.

In our setting we find that some alternative hypothetical choice protocols reduce the overall degree of hypothetical bias, but they appear to do so by introducing offsetting biases, rather than by addressing the underlying cause of the bias. We consider hypothetical choices elicited with the cheap talk script, as well as own and vicarious purchase likelihoods assessed on a 5-point scale, which we transform into binary choices by counting only the highest value ("very likely to purchase") as a hypothetical purchase.[34] For completeness, we also show results based on a binary transformation of the hypothetical WTP variable (labeled WTP choice), which infers a hypothetical intent to purchase item j at price p_j for individual i if $\text{WTP}_{ij} \geq p_j$.

As shown in Columns (4)–(7), two of the four alternatives magnify the bias, and a third yields only a modest improvement. The fourth alternative, a dichotomized vicarious choice, produces an estimate of -0.09, which is closer to the true effect. However, had we not known the ground truth, we would have had no basis for selecting the dichotomization threshold used for this estimate over other thresholds, which yield less accurate estimates. Moreover, it appears that the improvement is accidental, and does not reflect more informative responses. In particular, the lower half of the table reports correlations between real choices and the various hypothetical measures, both in levels (at a given price) and differences (changes

similar conclusions.

[34] Using other thresholds leads to worse estimates of the treatment effect.

Table 2: Snack Demand Treatment Effects: Univariate Specifications

	Ground Truth	Observational	Hypothetical as Prediction					Hypothetical as Predictors				
	Experiment	Diff. in Outcomes	Diff. in Hypotheticals					Low Dimensional				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Median estimated effect of high price	-0.075	-0.025	-0.159	-0.188	-0.129	-0.091	-0.266	-0.079	-0.083	-0.063	-0.047	-0.091
Median standard error	(0.004)	(0.014)	(0.006)	(0.007)	(0.006)	(0.005)	(0.009)	(0.008)	(0.010)	(0.009)	(0.006)	(0.012)
Hypotheticals:												
...hypothetical choice			X					X				
...cheap talk				X					X			
...intensity as choice					X					X		
...vicarious as choice						X					X	
...WTP as choice							X					X
Sample size (outcome)	189 ($\times 2$)	189	189	189	189	189	189	189	189	189	189	189
Univariate correlation with truth												
...levels	1.00	-	0.75	0.69	0.64	0.64	0.60	-	-	-	-	-
...difference	1.00	-	0.44	0.42	0.18	0.25	0.14	-	-	-	-	-
Observed at high price	All	$\text{WTP}_j > \epsilon_{jr}$	irrelevant					$\text{WTP}_j > \epsilon_{jr}$				
Observed at low price	All	$\text{WTP}_j \leq \epsilon_{jr}$	irrelevant					$\text{WTP}_j \leq \epsilon_{jr}$				

Estimates of the effect of the high price (vs. low price) on the real purchase frequency. Treatment is assigned endogenously based on the continuous average WTP variable. The reported estimates and standard errors are the median values across 10,001 simulated samples, which only differ by treatment assignment and hence observed outcome.

between high and low prices). The overall correlation between real demand and the standard-protocol hypothetical 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. In particular, the correlation between vicarious choices and real outcomes is noticeably lower than for the standard protocol (0.64 versus 0.75 in levels, 0.25 versus 0.44 in differences). This result could reflect a tendency to respond more randomly to vicarious questions, which would attenuate the difference between the means at different prices. However, all of these hypothetical responses are clearly correlated with real choices, and thus may make useful predictors. It seems likely that different protocols elicit different (and potentially complementary) information.

In contrast, by using hypothetical responses as predictors, our method largely removes the bias resulting from treatment endogeneity, even when hypothetical choices are systematically biased. In the final columns of Table 2, we exhibit estimators based on univariate models that relate the outcome to each hypothetical variable individually. For the estimators that use standard hypothetical choices, cheap-talk responses, or own-choice likelihoods, the estimates range from -0.063 to -0.083. The estimator that uses vicarious-choice likelihoods is a bit less accurate (-0.047), but is still in the ballpark. For completeness, we also include an estimator that uses the dichotomized WTP choice, even though the exercise presupposes that the WTP data are not available. Overall, using even a single hypothetical choice variable as a predictor rather than as a prediction shows promise for removing bias arising from treatment endogeneity.

Our method may perform even better when it employs multiple hypothetical covariates

that more comprehensively span motivations. Table 3 explores this possibility. Column (1) reproduces the true average treatment effect. The next two columns investigate whether standard methods yield more accurate estimates of treatment effects when they include controls for conventional covariates (physical characteristics, including grams per serving and seven measures of nutrients) in a regression of the outcome on the treatment. Column (2) reports an OLS regression. To allow for nonlinearities, we also use approximate residual balancing (ARB, Athey et al., 2018) with the same covariates as well as second-order terms and interactions (Column (3)).[35] For our method, we show results based on several specifications of the prediction model. For Column (4), we use all four hypothetical choice variables together (but exclude WTP, which governs treatment assignment). For Column (5), we add the eight physical characteristics. For both of these versions, we estimate the prediction model using OLS. We also consider three high-dimensional specifications, for which we use ARB as described in Section D.2. The first of these (Column (6)) includes the four hypothetical choice variables and eight physical characteristics, as well as second order and interaction terms. The second specification (Column (7)) uses more detailed information concerning the distributions of responses to the hypothetical choice elicitations, as well as other types of hypothetical reactions that potentially capture disaggregated motivations such as health concerns (we list the covariates in Appendix E.2). The third specification (Column (8)) adds a complete set of second-order and interaction terms.

Controlling for conventional covariates in a regression of the outcome on the treatment (Columns (2) and (3)) yields estimates in the neighborhood of -0.03, which is closer to the raw differences in means reported in Column (2) of Table 2 (-0.025) than to the true effect (-0.075). In contrast, the multiple-covariate versions of our method yield estimates between -0.071 and -0.081. The most accurate specifications (Columns (5) and (6)) include the four basic hypothetical choice variables and condition on physical characteristics.

3.3 Effect of an unseen counterfactual

Our method can also reveal treatment effects in applications for which there is no real-world variation in the treatment of interest. Such environments trivially satisfy unconfoundedness (Assumption 4), but the reliability of the estimate depends on the accuracy with which the relationship between choices and hypothetical responses extrapolates into the unseen setting.[36]

[35] Estimates using other doubly robust methods, such as those of Chernozhukov et al. (2018), yield similar results.

[36] Theoretically, extrapolation is accurate when the mapping from predictors to outcomes is invariant (Assumption 2), as long as either the distributions of evaluations for the hypothetical treatment states are overlapping (Assumption 5) or the relationship is linear (Assumption 3).

Table 3: Snack Demand Treatment Effects: Multivariate and High-Dimensional Specifications

	Ground Truth	Observational		Hypotheticals as Predictors				
	Experiment	OLS	ARB	Low Dimensional		High Dimensional		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Median estimated effect of high price	-0.075	-0.030	-0.028	-0.081	-0.077	-0.075	-0.081	-0.071
Median standard error	(0.004)	(0.014)	(0.013)	(0.009)	(0.008)	(0.008)	(0.005)	(0.011)
Controls		X	X		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		X
Sample size (outcome)	189 (×2)	189	189	189	189	189	189	189
Observed at high price	All	$\text{WTP}_j > \epsilon_{jr}$		$\text{WTP}_j > \epsilon_{jr}$				
Observed at low price	All	$\text{WTP}_j \leq \epsilon_{jr}$		$\text{WTP}_j \leq \epsilon_{jr}$				

Estimates of the effect of the high price (vs. low price) on the real purchase frequency. Treatment is assigned endogenously based on average WTP. The reported estimates and standard errors are the median values across 10,001 simulated samples, which only differ by treatment assignment and hence observed outcome.

3.3.1 Results

Table 4 shows that even if we observe all snacks at the high price (top panel) or all snacks at the low price (bottom panel), we can obtain reasonable estimates of the treatment effect. Column (1) reproduces the true average treatment effect, while the rest of the columns employ variants of our method. The first two variants use univariate prediction models: for Column (2), the predictor is the standard hypothetical choice, while for Column (3) it is the dichotomized WTP choice (recall that simulated treatment assignment is not governed by WTP in these simulations).[37] When all snacks are observed at the high price, both specifications yield estimates close to the true average effect. However, when all snacks are observed at the low price, the specification using WTP choice is considerably less accurate. Below, we show that this instability may be traceable to a violation of our assumption concerning overlapping evaluations. Columns (4)–(8) employ specifications analogous to those in Table 3, except that here we include dichotomized WTP responses throughout. The estimates are close to the true average effect when all snacks are observed at the high price. There is less stability when all snacks are observed at the low price, in that two of the three estimates are noticeably farther from the truth.

3.3.2 Discussion

When predicting the outcome for an unseen treatment state, our method projects from the space of treatments, where variation is absent, into the space of motivations, where

[37] Estimates for the other univariate specifications in Table 1 are in the Online Appendix.

Table 4: Estimating Treatment Effects without Variation in Treatment

	Ground Truth	Our Method: Hypotheticals as Predictors						
	Experiment	Low Dimensional				High Dimensional		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Observing all snacks at high price								
Estimated effect of high price	-0.075	-0.082	-0.078	-0.084	-0.077	-0.085	-0.093	-0.090
standard error	(0.004)	(0.008)	(0.013)	(0.011)	(0.011)	(0.016)	(0.005)	(0.014)
	[0.004]	[0.007]	[0.011]	[0.010]	[0.010]	[0.020]	[0.021]	[0.025]
Observed at high price	All				All			
Observed at low price	All				None			
Observing all snacks at low price								
Estimated effect of high price	-0.075	-0.084	-0.147	-0.119	-0.116	-0.140	-0.131	-0.073
standard error	(0.004)	(0.008)	(0.016)	(0.013)	(0.014)	(0.015)	(0.006)	(0.031)
	[0.004]	[0.006]	[0.014]	[0.013]	[0.014]	[0.019]	[0.025]	[0.028]
Observed at high price	All				None			
Observed at low price	All				All			
Controls					X	X	X	X
Hypotheticals:								
. . . hypothetical choice		X		X	X	X	X	X
. . . WTP as choice			X	X	X	X	X	X
. . . all hypothetical choices				X	X	X	X	X
. . . detailed hypothetical eval.							X	X
2nd order + interactions						X		X
Sample size (outcome)	189 ($\times 2$)	189	189	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.

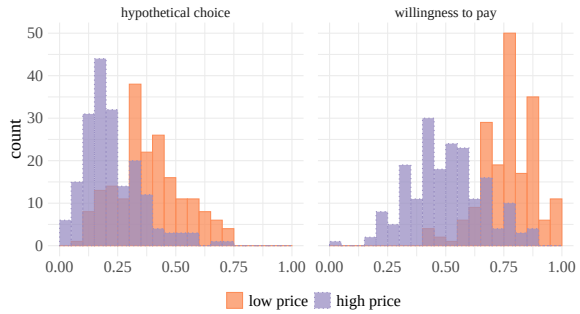
characteristics vary over the same dimensions irrespective of treatment state. This feature makes extrapolation feasible.

Figure 2 shows what can go wrong when overlap is incomplete. Part (a) depicts the distributions of evaluations for the high price in purple and for the low price in orange. The analyst can generate such overlap plots in any application, even without observing ground truth. The upper left panel focuses on the standard hypothetical choice variable. The distribution of this variable with the low price fully spans the distribution with the high price, and vice versa. When we estimate the relationship between hypothetical choice and outcomes based on all snacks at one price, this mutual spanning property allows that relationship to accurately predict outcomes at the other price (see column (2), both panels). In contrast, spanning for the WTP choice variable is asymmetric, as shown in the upper right panel. While the distribution of the WTP choice with the high price spans the distribution at the low price, the opposite is not true: there are very few snacks for which fewer than half the respondents report a hypothetical WTP below the low price of \$0.25. As a result, if we were to observe all real choices at the low price, predicting purchases at the high price based on WTP choice would require extrapolation beyond the range of observation. Hence, for the WTP choice, we can predict more confidently from high price to low price than in the opposite direction.

Part (b) uses observations of actual demand at both prices to show that the predictive relationship may be approximately linear for one measure (standard hypothetical choice) but not for another (WTP choice, which exhibits nonlinearity at lower values). In practice, if we observed all snacks at the low price, we would only be able to plot the orange squares, from which we might infer the orange curves. Because the low-price data does not span hypothetical WTP purchase frequencies far below 0.5, it cannot reveal that the relationship becomes markedly non-linear over that range. We can uncover this property in our experiment (for which we actually have real choices at both prices) by inspecting the high-price data (the purple curve).

As this example illustrates, when our method does not have access to real choices in the unseen treatment state, it relies heavily on the assumptions of either overlap (Assumption 5) or linearity (Assumption 3). When using our method for unseen counterfactuals, care should be taken to inspect and evaluate overlap.[38] Hypothetical data are more likely to satisfy overlap when the variation from the treatment is small relative to that arising from other factors.

[38]While one can also check linearity, our example strikes a cautionary note: a relationship may be linear only within the observed (overlapping) range of variation.



(a) Overlap in Hypotheticals

(b) Relationship between Outcome and Hypothetical

Figure 2: Overlap between Hypothetical Evaluations

3.4 Heterogeneity in treatment effects

When hypothetical evaluations are highly predictive of outcomes, they may also reveal heterogeneity in treatment effects that is difficult to quantify using standard methods. In this section, we compare the performance of various methods for measuring heterogeneous treatment effects, and examine implications for optimal price setting.

3.4.1 Metrics

We report four measures of the degree to which the estimated heterogeneity in treatment effects, $\hat{\tau}_j = \hat{Y}_j(1) - \hat{Y}_j(0)$ for unit j, captures the heterogeneity in actual effects, $\tau_j = Y_j(1) - Y_j(0)$:

- R^2 for a regression of τ_j on $\hat{\tau}_j$. This statistic measures the fraction of the variation in true treatment effects that the estimates capture.
- *Mean squared error* ($\text{mse} := \text{mean}((\tau_j - \hat{\tau}_j)^2)$)*:* This statistic encompasses overall accuracy and precision.
- *Calibration coefficient:* This measure is the slope coefficient in a regression of τ_j on $\hat{\tau}_j$. The ideal coefficient is unity: in that case, the expectation of the actual treatment

effect increases unit for unit with the predicted treatment effect.[39]

- *Simulated profit:* We simulate a producer who estimates heterogeneous price sensitivity for each snack j in order to set prices w_j^*. We report the gain in average profit, relative to setting prices at random, as a fraction of the maximum possible gain achieved by optimal pricing, $\frac{\bar{\pi}(\mathbf{w}^*)-\bar{\pi}(\mathbf{w}^{random})}{\bar{\pi}(\mathbf{w}^{optimal})-\bar{\pi}(\mathbf{w}^{random})}$. For this calculation, we define profit as $\pi_j(w) = (w \cdot 0.75 + (1-w) \cdot 0.25 - c)Y_j(w)$ for $w \in \{0,1\}$ for snack j and average profit as $\bar{\pi}(\mathbf{w}^*) = \frac{1}{J}\sum_j \left[(w_j^* - c)Y_j(w_j^*)\right]$. We set marginal costs c so that it is optimal to sell half of the snacks at the low price and half at the high price.[40] The producer observes demand for snack j at a single price W_j, predicts demand at the other price, $\hat{Y}_j(1-W_j) = Y_j(W_j) + \hat{\tau}_j \cdot (1_{\{w>W_j\}} - 1_{\{w<W_j\}})$, and sets the price to maximize predicted demand: $w_j^* = \arg\max_w (w \cdot 0.75 + (1-w) \cdot 0.25 - c) \cdot \hat{Y}_j(w)$. [41] The producer achieves optimal profits when $\hat{\tau}_j = \tau_j$ for all j; imperfect estimates result in lower profits.

3.4.2 Results

Results appear in Figure 3. Until indicated otherwise, we abstract from endogeneity and focus on environments with random treatment assignment, which we simulate by selecting half of the snacks (94 of 189) at random to serve as the treated units. For each estimation method, we plot each metric's median value and interquartile range based on 10,001 simulated samples.

Row 1 corresponds to the difference-in-means estimator, $\hat{\tau}_j \equiv \hat{\tau} = \frac{1}{\sum_{j'=1}^J W_{j'}} \sum_{j'=1}^J W_{j'} Y_{j'} - \frac{1}{\sum_{j'=1}^J 1-W_{j'}} \sum_{j'=1}^J (1-W_{j'}) Y_{j'}$, which we offer as a simple benchmark. Because this estimate does not vary with j, R^2 and the calibration parameter are both zero. Even so, if the available covariates have little explanatory power, this simple estimator may perform well in terms of MSE and simulated profits by virtue of its parsimony.

[39] Typically, there is some trade-off between the calibration coefficient and R^2. For instance, one can increase the calibration coefficient by projecting predicted effects onto a binary covariate. Because this procedure reduces the noise in the predicted treatment effects, it tends to increase the calibration coefficient. At the same time, the projection removes some of the signal along with the noise, which reduces R^2. The calibration coefficient is also a measure of the excess variation of treatment effect estimates. To understand why this is the case, suppose the estimated treatment effect is an unbiased estimate of the actual treatment effect: $\hat{\tau}_j = \tau_j + \epsilon_j$, where ϵ_j is mean zero and independent of τ_j. Then, by standard calculations for classical measurement error in regressors, the calibration coefficient is $\frac{\text{var}(\tau_j)}{\text{var}(\tau_j)+\text{var}(\epsilon_j)} \leq 1$.

[40] Because the real demand response to tripling prices is relatively small for most snacks, this procedure yields a negative value of marginal cost ($c = -1.25$). For this value, 86 (out of 189) snacks are more profitable at the high price, 91 are more profitable at the low price, and 12 are equally profitable at the two prices. While a negative marginal cost is obviously implausible, the point of the simulation is simply to show how more accurate estimates of heterogeneous responses can impact optimization.

[41] We focus on this rule because it is simple and plausible; other pricing policies may perform better according to some metrics because the producer estimates optimal prices with variance.

Conventional estimators identify heterogeneous effects by conditioning on a set of observed characteristics. For row 2, we linearly project the actual unit-level treatment effect on all the physical characteristics. Because this approach requires us to observe each unit in both treatment states, it is infeasible under the assumptions governing this exercise. However, it provides a useful benchmark because it quantifies the greatest amount of heterogeneity one might hope to capture through this conditioning approach.[42] We also consider three conventional estimators that are feasible in the sense that they only use data for one treatment state per unit: separate OLS estimates, by treatment status, of linear relationships between the outcome and all physical characteristics; a similar LASSO approach that adds interactions and second-order terms; and a causal forest (Wager and Athey, 2018) with the eight physical characteristics as features.

Our method captures substantially more unit-specific heterogeneity beyond that associated with the physical characteristics. Row 6 of Figure 3 shows results for the variant that employs hypothetical choices and physical characteristics as predictors (i.e., the same variant as in Table 4 Column (5)). Performance measures are substantially better across the board compared with the three feasible conventional estimators. Our method also easily surpasses the infeasible benchmark with respect to all metrics other than calibration. The latter comparisons imply that hypothetical evaluations contain substantially more information about variation in treatment effects than physical characteristics in our setting.

Having shown that our method can capture substantially more treatment effect heterogeneity at the unit-level than conventional methods, we next compare performance when we use our method only to extract heterogeneity that is related to the same observable characteristics. For this purpose, we linearly project the estimated treatment effects onto the physical characteristics. The resulting estimates (row 7) generally perform as well as or better than the feasible conventional methods. The improvements reflect the fact that our unit-level estimated effects contain information (from hypothetical evaluations) about both treatment states for each snack, and we use all of that information when projecting onto physical characteristics.

So far, we have focused on simulations with randomized treatment assignment. When treatment is assigned endogeneously, existing quasiexperimental methods rely on the component of variation that is plausibly exogenous (the variation arising from compliers) to estimate treatment effects. As a result, those methods may not be sufficiently precise to detect much heterogeneity, or if they are, the heterogeneity they detect may result from systematic differences in the sets of compliers between settings, rather than actual differences in treatment effects. In contrast, our method still performs well in terms of recovering

[42] In the figure, the interquartile ranges are degenerate because the results do not depend on the simulated treatment assignments.

heterogeneous effects according to all four metrics. For row 8 of Figure 3, we use the same prediction model as in row 6, but we apply our method to simulated draws based on the endogenous assignment rule described Section 3.2.1. Compared to the environment with random treatment assignment (row 6), we find only modest deterioration of the method's performance, which is presumably attributable to the small bias associated with the estimate in Column (5) of Table 4. Our method noticeably outperforms feasible conventional approaches that condition on physical characteristics (rows 3-5) even when we handicap it (and not the alternative methods) by introducing endogenous treatment selection.

Because our method does not require an intervention, it can enable analysts to recover heterogeneous treatment effects even when they lack the power to intervene. This feature may be particularly valuable in settings where one wishes to target the treatment at those who would benefit most.

3.5 Gains in precision

Our method may also yield more precise estimates of treatment effects than conventional alternatives even when experimental evidence is available. Most notably, the performance of standard methods deteriorates when the fraction treated is far from half, while our method maintains good performance even if few of the observations are treated (or none, as in Section 3.3). It may be far cheaper and more convenient in practice to reduce variance by collecting hypothetical responses, rather than by expanding the experimental sample.

We explore these issues in an environment with random treatment assignment (no endogeneity). Fixing the fraction of snacks observed at the higher price, we simulate uncertainty in treatment assignment by randomly dividing the snacks into high-price and low-price subsets of fixed sizes. We generate 10,001 such random samples. We then compute the standard deviation, bias, and root-mean-squared error for various treatment effect estimators. These metrics hold fixed the snacks that are in the sample, their covariates (physical characteristics and hypothetical evaluations), and their outcomes for each treatment state.

We consider two standard approaches, difference-in-means and the ARB estimator from Column (3) of Table 3, as well as two variants of our method, the univariate specification using the standard hypothetical choice and the high dimensional specification from Column (8) of Table 4.[43] Figure 4 plots the resulting statistics as functions of the fraction of snacks observed at the high price.

The standard deviations of our estimators are substantially smaller than those of the conventional estimators, especially when the proportion treated is far from half, as shown

[43] For results based on all specifications of our method from Tables 2 and 3, see the Online Appendix.

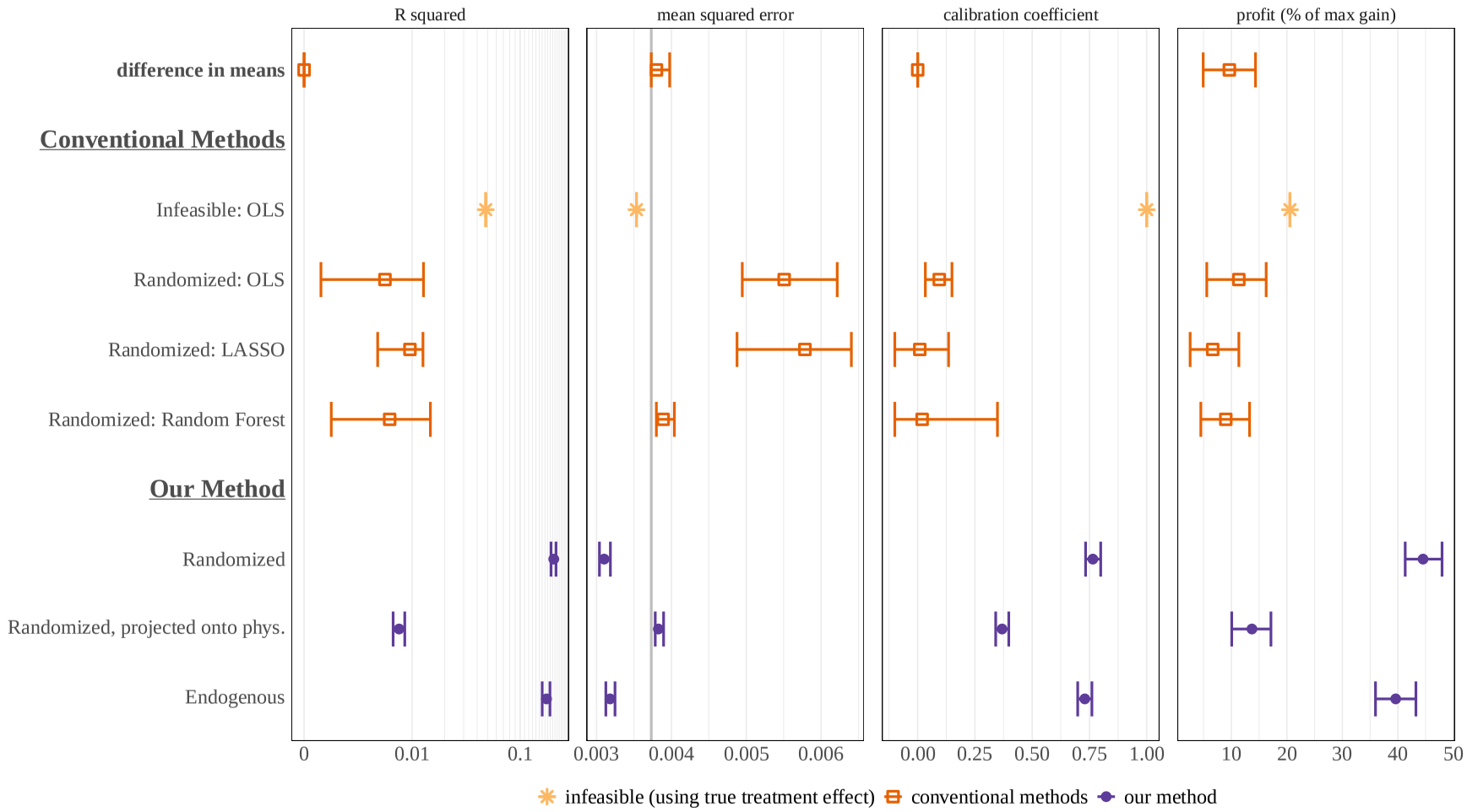


Figure 3: Treatment Effect Heterogeneity

Summary statistics describing how well different estimators capture heterogeneity in treatment effects. Points indicate the median value of each statistic across 10,001 simulated samples, and whiskers indicate the interquartile range. The R-squared axis is an augmented log scale that includes 0 where 0.001 would be on a regular log scale. For mean squared error, the vertical line shows the value obtained when we use the true average treatment effect without any heterogeneity; in other words, it is the variance of unit-level treatment effects. For the calibration coefficient, the lower boundary of the first quartile for the random forest estimator is -0.857, but is shown in the figure as -0.1 because the axis is truncated.

in the first panel of Figure 4.[44] The standard deviation of the difference-in-means is U-shaped in the fraction of treated observations. When half of the sample is treated, the (median) standard error of the difference-in-means is more than twice that of the univariate hypothetical choice estimator.[45] To achieve the same standard error for the difference-in-means as for our univariate hypothetical choice specification with 189 snacks, one would need a randomized experiment with over 800 snacks. As the sizes of treated and untreated subsamples become more unbalanced, conventional estimators lose dramatically more precision because the smaller of the treatment and control groups dominates the variance. In contrast, the precision of our low-dimensional estimator is largely independent of the proportion treated. The reason is that the first step of our method pools all observations, and the second uses the hypothetical evaluations for *both* treatment states for every snack.[46]

In this application, a smaller standard deviation comes at the cost of a small bias (Figure 4 second panel), but our estimators attain lower root-mean-squared error, irrespective of the treatment's prevalence (Figure 4 final panel). The difference-in-means is unbiased by design, and hence its root-mean-squared error is equal to its standard deviation. The univariate hypothetical choice method entails a slightly larger bias, but the reduction in variance more than compensates in terms of root-mean-squared error.[47] 95% confidence intervals for our estimators achieve their nominal coverage of the true treatment effect in these simulations, as we show in Appendix Figure A4.

4 Application: Microfinance Contributions

To boost fundraising, non-profit organizations often inform potential contributors that other donors have agreed to match contributions (Dove, 1999; List, 2011). How well does this

[44] While these results pertain to a fixed sample of snacks, the standard error formulas also reflect the additional variation associated with sampling snacks (independently) from some super-population. For the difference-in-means estimator, the *sampling-based* variance exceeds the *design-based* variance by the variance of treatment effects divided by sample size. The exact, design-based, finite sample variance of the difference-in-means estimator in these simulations is $\text{var}(Y_j(1))/J_1 + \text{var}(Y_j(0))/J_0 - \text{var}(\tau_j)/J$ (cf. Imbens and Rubin, 2015). When sampling from an infinite super-population, the variance of treatment effects, $\text{var}(\tau_j)/J$, is dropped from the formula. In Appendix Figure A4, we show that the computed standard errors of the regression estimators similarly overstate their design-based variances in these simulations.

[45] Figure 4 presents simulated standard deviations; for estimated standard errors and probability of coverage of confidence intervals, see Appendix Figure A4.

[46] The high-dimensional variant of our approach also yields greater precision than the standard methods, but the gains are not as dramatic for imbalanced samples. The associated standard deviation is U-shaped because, with extreme imbalance, evaluations overlap tends to be poor, and residual balancing attributes greater weight to the few observations that do provide overlap.

[47] The standard ARB estimator introduces a small finite-sample bias, and does not reduce variance sufficiently to achieve an overall reduction in root-mean-squared error for this application. In contrast, the high-dimensional version of our method reduces bias and consequently performs comparably to the univariate version in terms of root-mean-squared error when the fraction of snacks observed at the high price is close to one half. For less well-balanced designs, the marked difference in standard errors overwhelms the difference in bias, causing the univariate specification to perform unambiguously better in terms of mean squared error.

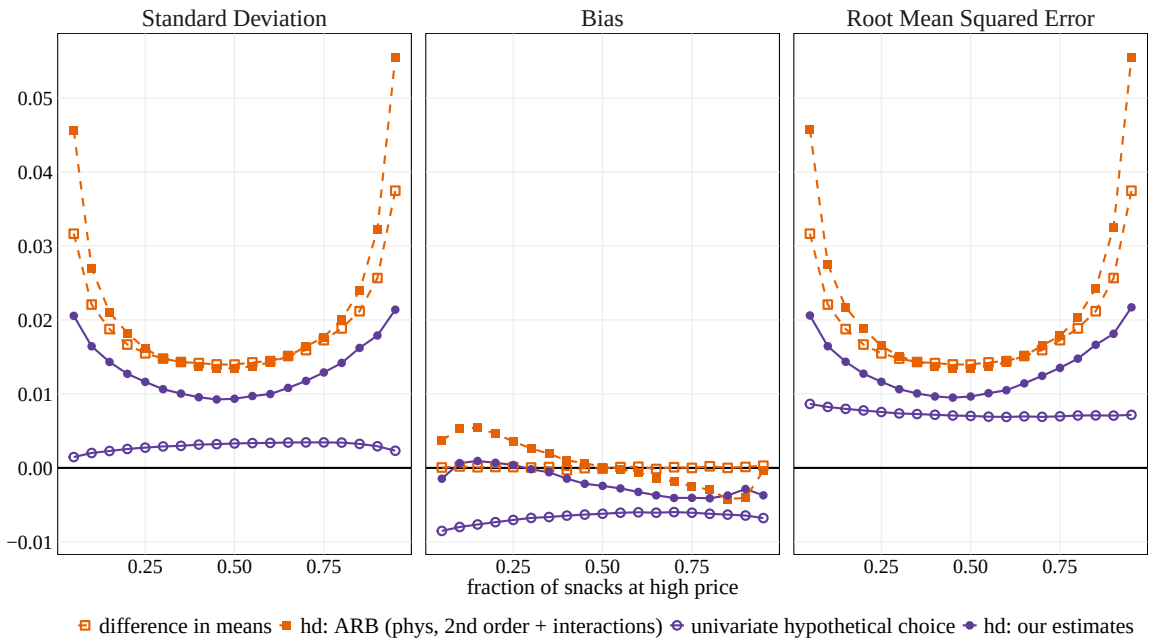


Figure 4: Performance of Estimators by Fraction Treated

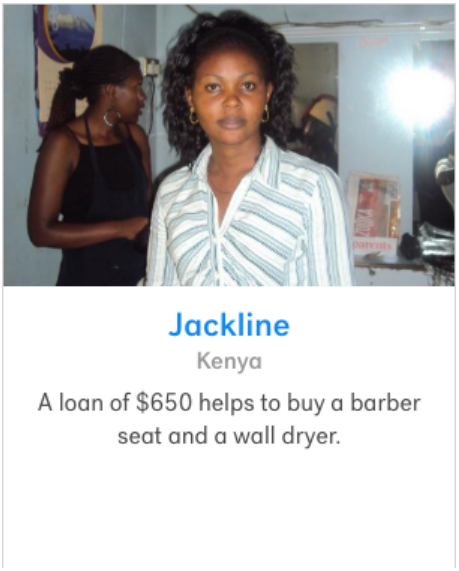
Summary statistics describing properties of treatment effect estimators under random assignment. The horizontal axis measures the fraction of snacks observed at the high price. At the boundaries of the interval, only our estimators are well-defined (see also Section 3.3), and the standard deviation (across realizations of the assignment distribution) is mechanically zero because there is only one possible assignment.

strategy work? Estimating the causal effects of a match is challenging when the match is not randomly assigned (Karlan and List, 2007; Huck and Rasul, 2011). In this section, we use our method to determine the impact of matching provisions in the context of microfinance.

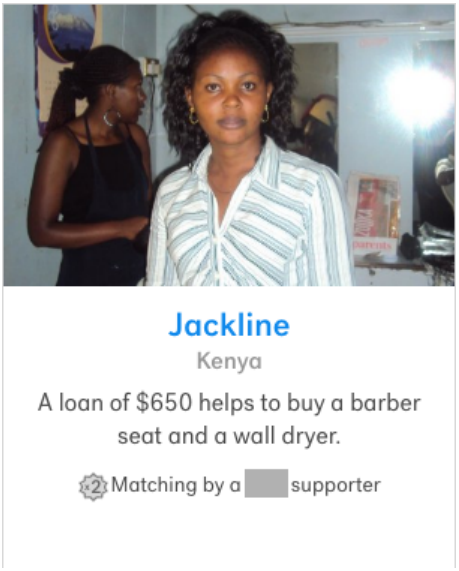
We focus on a large microfinance crowdsourcing website, which displays profiles of potential borrowers and allows website visitors to contribute to their loans. Contributors are typically socially minded individuals in developed countries; borrowers are typically developing country residents who request funds for various projects (such as business, agricultural, home, or health expenses). Sponsors selectively offer matching funds. When a loan is eligible for a match, the profile prominently displays an indicator (as shown on Figure 5). For every dollar the visitor contributes, the sponsor also contributes a dollar.

Assignment of the treatment (matching) is complex and endogenous. The website cultivates sponsors to provide funds for matching loans. Sponsors can 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). They may be individuals or collectives, such as churches or community groups. Endogeneity arises from correlations between the preferences of sponsors and contributors. Conditioning the analysis on the potential matching criteria does not resolve this issue.[48]

[48] A fully conditional model would be extremely high-dimensional, in that it would control for all combinations



(a) Unmatched profile



(b) Matched profile

Figure 5: Loan Profiles with Matching Indicator

For this application, the treatment unit j is a loan profile, and the treatment $w \in \{0, 1\}$ specifies whether the loan is matched. The 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.[49] The treatment effect of interest is the average impact of matching on fundraising.

We compare estimates from observational data using standard methods and our method, against a ground truth estimate based on a field experiment in which we introduce randomly assigned matches.[50]

4.1 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.

of criteria sponsors are allowed to specify. Furthermore, there is no reason to think treatments would be exogenous in such a model. Sponsors may decide to match certain types of proposals based on transient factors that may also influence contributions, such as news stories that draw attention to particular countries, or the attractiveness of postings within particular categories at the time of the matching decision.

[49]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; see, for instance, Bellemare and Wichman (2020) regarding its interpretation and use in economics.

[50]This experiment was preregistered (AEARCTR-0004885).

4.1.1 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 (we omit 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 (because they were matched for at least 90% of the first 24 hours after their initial posting) or unmatched (because they were matched for no more than 3% of the first 24 hours). We drop the remaining 18% of profiles, which were matched for intermediate fractions of the first 24 hours, to create a binary treatment indicator.[51] According to this criterion, 623 (6.5%) of the retained profiles were matched. For each of these profiles, our data include descriptive characteristics, when it was matched, and how quickly it raised funds.

4.1.2 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, oversampling matched loans. 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. They then viewed a random draw of 30 loan profiles from the set of 300, each of which appeared either in the same treatment state as on the website, or edited to add or remove the matching funds indicator. We displayed 20 of the 30 loans as unmatched (10 of which were actually unmatched on the website) and 10 as matched (5 of which were actually matched on the website).

Participants rated each (real or counterfactual) loan profile by predicting a quintile for fundraising velocity on the first day, indicating the likelihood they would lend $25 to it, and indicating the likelihood a typical user would lend $25 to it (both 7-point Likert 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 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 evaluations of 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 F.1.

[51] Observational methods yield similar estimates of the treatment effect when we retain all profiles and control linearly for the fraction of time each profile was matched during the first 24 hours.

4.1.3 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%).[52] 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 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 sponsorship account with positive balances. Loans that would be matched irrespective of our intervention are always-takers; they correspond to matched loans in the observational sample. The effect on always-takers corresponds to the average treatment effect on the treated (ATT). Loans that would not have been matched in the absence of our intervention, whether in the control or treatment group, are compliers. The population of compliers in the experiment corresponds to unmatched loans in the observational data, and the local average treatment effect (LATE) corresponds to the average treatment effect on the control (ATC).[53]

4.2 Local Average Treatment Effects

Table 5 contains estimated treatment effects for matching provisions (τ for control loans, LATE) derived through a variety of methods. For the experimental sample, the assignment of matching is random for compliers, so we use the standard instrumental variables estimator. We estimate a treatment effect of 1.24 (s.e. 0.33), which we treat as ground truth.

Next we attempt to recover treatment effects using only the observational data. As we have noted, it is difficult to unravel the structure of the process that renders matching provisions endogenous, and good instruments are difficult to find. 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 the ground truth. Adding standard controls does not help: whether we insert each factor linearly (Column (3)) or flexibly control for linear, quadratic, and interaction terms using ARB (Column (4)), the estimate drifts further from the truth. We reject equality between each of these estimates

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

[53] 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 (cf. Angrist et al., 1996).

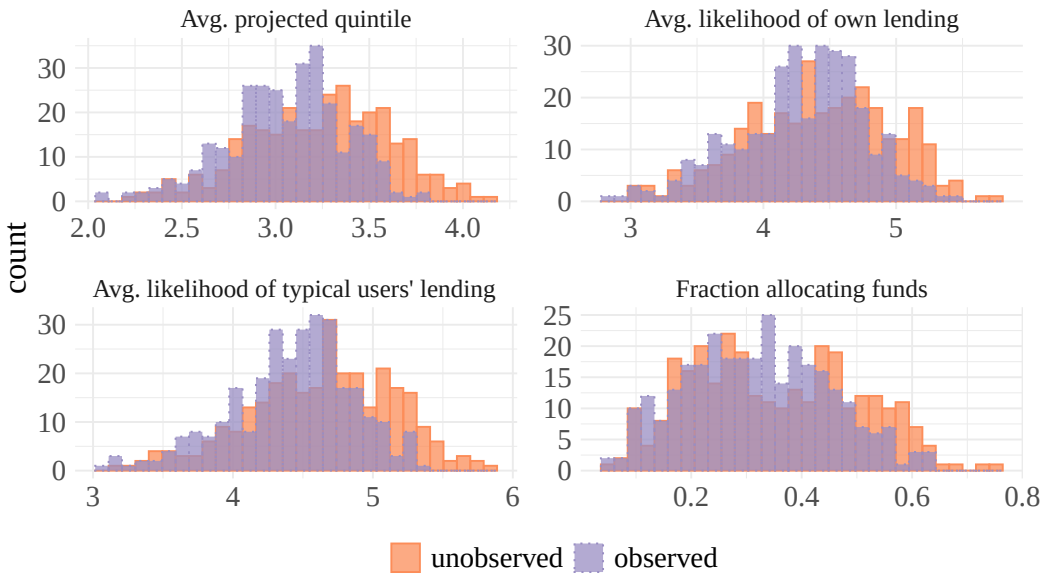


Figure 6: Overlap in hypothetical evaluations for loan × treatment states that are observed (blue) vs. unobserved (red) in the data.

and the ground truth.

Next we turn to estimates based on hypothetical evaluations. We begin by checking overlap – that is, whether the distribution of evaluations over profiles in the observed treatment states span the corresponding distribution in the unobserved treatment states. Figure 6 shows that, for most of the evaluations of profiles in unobserved states (red), there are indeed loans with similar evaluations in their observed states (blue). Consequently, our method requires only modest extrapolation (for high desirability).

Our method yields estimates that are close to the ground truth results. Table 5 exhibits a low-dimensional specification that includes the average of each hypothetical evaluation (Column (5)) and one that adds standard controls (Column (6)), as well as high-dimensional specifications estimated with ARB that add quadratic and interaction terms (Column (7)), distribution detail for each possible hypothetical response (Column (8)), and both (Column (9)). Estimates range from 0.90 to 1.63, and statistical tests fail to reject the hypothesis that each coincides with the ground truth.

4.3 Heterogeneity: Treatment Effects by Complier Group

The instrumental variables procedure yields estimates of the treatment's effect on compliers (a LATE). This focus is a limitation of experimental and quasiexperimental approaches (see, for instance, Deaton, 2010; Heckman and Urzúa, 2010; Imbens, 2010, for a discussion). In many applications, the analyst may be interested in treatment effects for other groups. For example, if we were interested in the effects of eliminating the microfinance website's

Table 5: 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)
Estimated effect of matching	1.24	2.55	3.21	3.10	0.90	1.04	1.63	1.01	1.39
Analytical standard error	(0.33)	(0.33)	(0.30)	(0.37)	(0.25)	(0.24)	(0.25)	(0.18)	(0.25)
Bootstrap standard error	[0.32]	[0.34]	[0.30]	[0.29]	[0.26]	[0.24]	[0.35]	[0.30]	[0.42]
Test: = ground truth (p-value)	1	0.01	0	0	0.42	0.62	0.41	0.60	0.77
Controls			X	X		X	X	X	X
Hypotheticals:									
... avg. hypothetical eval.					X	X	X	X	X
... freq. hypothetical eval.								X	X
2nd order + interactions				X			X		X
Sample size	1091	300	300	30	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.' incluces 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 + interactions' includes quadratic terms for the mean responses and frequencies of each hypothetical response, and all two-way interactions between mean responses, frequencies of each hypothetical response, and the controls. Analytical standard errors in parenthesis, bootstrap standard errors in square brackets.

Table 6: Heterogeneity by Compliance Group in the Microfinance Application

	Experiment	Our Method		Proportion
	IV (1)	Low Dimensional (5)	High Dimensional (9)	of Observational Sample
Estimated effect of matching				
..... on compliers (LATE/ATC)	1.24 (se 0.32)	0.90 (se 0.26)	1.39 (se 0.42)	93.5%
..... on always-takers (ATT)	cannot be estimated	0.23 (se 0.17)	0.69 (se 0.35)	6.5%
... average (ATE)	cannot be estimated	0.86 (se 0.25)	1.35 (se 0.39)	100%
Test: equal effects (p-value)	—	0	0.18	

The first row of estimates reproduces results from Table 5, columns (1), (5), and (9) (as indicated in the column headings). Standard errors in parenthesis are based on the bootstrap.

matching provisions, the most pertinent consideration would be the effects of matching on funding velocity for loans that are currently match-eligible (always-takers). Similarly, when choosing between making different matching policies universal, we would like to compare their overall effects (ATEs).

Our method can in principle estimate average treatment effects for any specified subgroup. We illustrate this feature in Table 6. The first row reproduces selected estimates of the LATE (also the ATC) from Table 5, including the IV estimate, as well as two measures obtained through our method (corresonding to the low and high dimensional specifications in, respectively, columns (5) and (9) of Table 5). Estimates of effects on always-takers (ATTs) appear in the second row, and estimates of overall effects (ATEs) appear in the third. Because IV cannot reveal either of these effects, the corresponding cells do not contain estimates. Policymakers relying on IV methods must hope that the LATE is representative of the effects on these other populations.

Our method reveals that treatment effects in fact differ substantially among compliance groups. The second row shows that our estimates of the average treatment effect on the treated (ATT) is less than half the LATE/ATC for both specifications. The accuracy with which our method reproduces ground truth for the LATE/ATC increases confidence that the estimate of the ATT is also reliable. Loans that are matched in practice apparently do not benefit as much from the match, presumably because they are sufficiently attractive in other dimensions to raise funds irrespective of matching. In this case, the estimated ATEs are close to the LATE/ATCs because the population of always-takers is relatively small (6.5% of the total). Nevertheless, our finding has an immediate policy implication: the microfinance platform may be able to raise more funds by inducing sponsors to match contributions to loans that are intrinsically less popular among the website's users.[54]

[54] By way of analogy to our analysis of optimal snack pricing in Section 3.4, one could in principle maximize the total impact of a fixed matching fund by devising a targeting system based on finer estimates of heterogeneous treatment effects.

4.4 Using the subset of hypothetical respondents who are most skilled

In some applications, the survey respondents answering hypothetical questions may differ noticeably from the people whose choices determine the real outcomes. Here, 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.[55]

One can focus on the hypothetical respondents best able to predict real outcomes to both diagnose their suitability and fine-tune their composition, as suggested in Section 2.5. We illustrate this procedure by filtering respondents based on correlations between their "quintile projections" and actual fundraising velocities for loan profiles displayed in their actual treatment states.[56] We address reduced sample sizes using an IV procedure. Figure 7 shows (in blue) how the IV estimates vary with the threshold r^*. Provided we filter out evaluations by the lowest quality respondents (those for whom responses are *negatively* correlated with outcomes), the estimates fall into the range of 1.25 to 1.6.

We consider two methods for choosing the threshold for response quality.. The first is to use the threshold that minimizes the out-of-sample mean squared error. The figure shows the threshold this criterion selects, $r^* = 0.16$, for which the estimate is 1.6.[57] Because estimates are highly correlated across thresholds, it can be difficult to select between them. We therefore consider a second method that is less sensitive to the choice of threshold: for each threshold r, we use an approach derived from ARB that balances residuals based on data for other thresholds. See Appendix D.2 for a detailed description of the algorithm. As the purple dots in Figure 7 show, the resulting estimates are much less sensitive to the choice of the threshold: all of them are between 1.27 and 1.37, only slightly higher than ground truth (1.24). The ARB estimate for the threshold that minimizes out-of-sample mean squared error for our IV procedure ($r^* = 0.16$) is 1.30.

5 Conclusion

In this paper, we have explored methods for inferring the causal effects of treatments on choices from data that include both real choices and hypothetical evaluations. We have

[55] 10% state they have made one loan using the website, and a little over 3% state they have made two or more loans. This issue does not arise in our snack application because we recruited the participants who make hypothetical choices from the same population as the participants who make real choices. For completeness we also show a corresponding analysis for the snack application in Appendix Figure A5 separately with all snacks observed at the high price and all snacks observed at the low price.

[56] For a respondent who gave the same answer concerning every loan, the correlation is undefined. We set it equal to -1, indicating the lowest possible response quality.

[57] For each threshold, we report the median across 11 random sample splits. See Appendix F.2 for details on our estimation of mean squared error with measurement error in regressors.

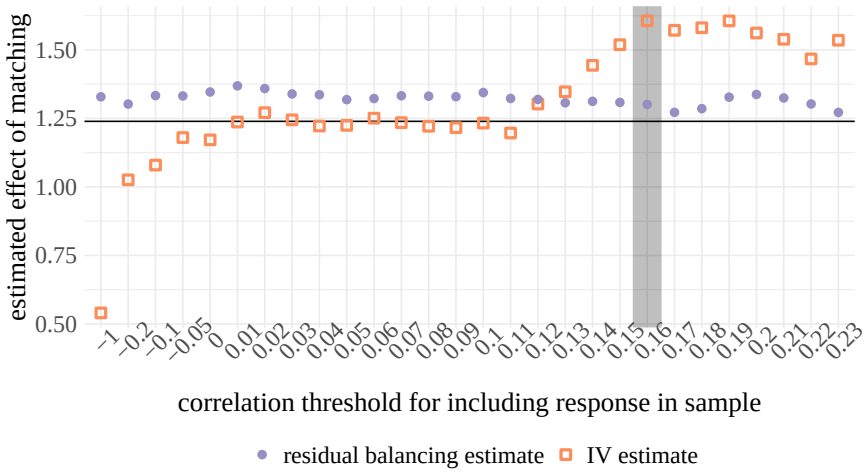


Figure 7: Estimates of the effect of matching by correlation threshold. Ground truth estimate shown as a horizontal line. Standard errors for each estimate are in the supplemental appendix.

proposed a class of estimators, identified conditions under which they yield consistent estimates, and derived their asymptotic distributions. In applications for which those conditions are plausible, the approach offers multiple advantages. First, it can recover average treatment effects in settings with endogeneity even when standard methods are inapplicable. Second, one can apply it even in cases for which there is no observed variation in the treatment (i.e., to evaluate an untested proposal). Third, it yields more comprehensive measures of heterogeneous treatment effects than standard approaches, in that it can recover treatment effects for arbitrary subgroups. Fourth, it can improve the precision of estimated treatment effects even when randomized variation is available, particularly when treatment groups are unbalanced. We have also provided proof of concept by applying the approach to data generated in a laboratory application, and through a field application on the effects of matching loan provisions offered on a large microlending platform.

We do not claim that the approach offers a panacea. Instead, we articulate the conditions under which it may be useful, and point to potentially suitable applications. Indeed, we do not recommend the method, as currently formulated, for certain classes of applications, such as those in which a single individual makes both the treatment selection decision and the outcome choice. That said, we anticipate that the approach will prove valuable in many settings. For example, it may provide a reasonably reliable and cost-effective alternative to field experiments, or it may complement field experiments by making it possible to explore large varieties of treatment possibilities before committing to a particular version.

An important unexplored question is whether the relationship between choices and basic motivations is stable, and therefore portable, over a broad domain. Our method only assumes

portability within a class of decision problems that may be relatively narrow for any particular application. 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 which case it may not be necessary to reestimate the relationship for each new application. Yet the hypothesized stable relationship may prove elusive due to the challenges associated with obtaining context-free measures of fundamental motivations. An interesting question, motivated by Smith et al. (2014), is whether neurobiological measurement can avoid contextual influences on reporting and capture the essence of those fundamental motivations more effectively than survey responses.

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." *Journal of the American Statistical Association* 105 (490): 493–505. 10.1198/jasa.2009.ap08746.

Abdellaoui, Mohammed, Carolina Barrios, and Peter P. Wakker. 2007. "Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory." *Journal of Econometrics* 138 (1): 356–378. 10.1016/j.jeconom.2006.05.025.

Ajzen, Icek, Thomas C. Brown, and Franklin Carvajal. 2004. "Explaining the Discrepancy between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation." *Personality and Social Psychology Bulletin* 30 (9): 1108–1121. 10.1177/0146167204264079, Publisher: SAGE Publications Inc.

Alpizar Rodriguez, Francisco, Fredrik Carlsson, and Peter Martinsson. 2003. "Using Choice Experiments for Non-Market Valuation." *Economic Issues Journal Articles* 8 (1): 83–110, https://econpapers.repec.org/article/eisarticl/103alpizar.htm, Publisher: Economic Issues.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–455. 10.1080/01621459.1996.10476902.

Athey, Susan, Raj Chetty, and Guido Imbens. 2020. "Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes." *arXiv:2006.09676 [econ, stat]*, http://arxiv.org/abs/2006.09676, arXiv: 2006.09676.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. 10.1257/jep.31.2.3.

Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (4): 597–623. 10.1111/rssb.12268.

Begg, C. B., and D. H. Y. Leung. 2000. "On the use of surrogate end points in randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163 (1): 15–28. 10.1111/1467-985X.00153.

Bellemare, Marc F., and Casey J. Wichman. 2020. "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bulletin of Economics and Statistics* 82 (1): 50–61. 10.1111/obes.12325.

Ben-Akiva, M., M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. 1994. "Combining revealed and stated preferences data." *Marketing Letters* 5 (4): 335–349. 10.1007/BF00999209.

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?" *American Economic Review* 102 (5): 2083–2110. 10.1257/aer.102.5.2083.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2014. "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference." *American Economic Review* 104 (9): 2698–2735. 10.1257/aer.104.9.2698.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112 (1): 68–105. 10.1086/379939.

Blackburn, McKinley, Glenn W. Harrison, and E. Elisabet Rutström. 1994. "Statistical Bias Functions and Informative Hypothetical Surveys." *American Journal of Agricultural Economics* 76 (5): 1084–1088. 10.2307/1243396, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/1243396.

Blamey, R. K., J. W. Bennett, and M. D. Morrison. 1999. "Yea-Saying in Contingent Valuation Surveys." *Land Economics* 75 (1): 126–141. 10.2307/3146997, Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johannesson, Nancy Horn, and Patricia Freeman. 2008. "Eliciting Willingness to Pay Without Bias: Evidence from a Field

Experiment." *The Economic Journal* 118 (525): 114–137. 10.1111/j.1468-0297.2007.02106.x.

Brandts, Jordi, and Gary Charness. 2009. "The Strategy versus the Direct-response Method: A Survey of Experimental Comparisons." *mimeo*, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.597.7870&rep=rep1&type=pdf.

Briggs, Joseph, Andrew Caplin, Søren Leth-Petersen, Christopher Tonetti, and Gianluca Violante. 2020. "Estimating Marginal Treatment Effects with Survey Instruments."

Brownstone, David, David S. Bunch, and Kenneth Train. 2000. "Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles." *Transportation Research Part B: Methodological* 34 (5): 315–338. 10.1016/S0191-2615(99)00031-4.

Carson, Richard T. 2012. "Contingent Valuation: A Practical Alternative When Prices Aren't Available." *Journal of Economic Perspectives* 26 (4): 27–42. 10.1257/jep.26.4.27.

Carson, Richard T., and Theodore Groves. 2007. "Incentive and informational properties of preference questions." *Environmental and resource economics* 37 (1): 181–210, Publisher: Springer.

Carson, Richard T., Theodore Groves, and John A. List. 2011. "Toward an Understanding of Valuing Non-Market Goods and Services." *mimeo, UCSD*.

Carson, Richard T., and W. Michael Hanemann. 2005. "Contingent valuation." *Handbook of environmental economics* 2 821–936, Publisher: Elsevier.

Champ, Patricia A., Richard C. Bishop, Thomas C. Brown, and Daniel W. McCollum. 1997. "Using Donation Mechanisms to Value Nonuse Benefits from Public Goods." *Journal of Environmental Economics and Management* 33 (2): 151–162. 10.1006/jeem.1997.0988.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21 (1): C1–C68. 10.1111/ectj.12097.

Conlon, Chris, Julie Mortimer, and Paul Sarkis. 2021. "Estimating Preferences and Substitution Patterns from Second Choice Data Alone." *Working Paper*.

Cummings, Ronald G., Glenn W. Harrison, and E. Elisabet Rutström. 1995. "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible?" *The American Economic Review* 85 (1): 260–266, https://www.jstor.org/stable/2118008, Publisher: American Economic Association.

Cummings, Ronald G, and Laura O Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method." *American Economic Review* 89 (3): 649–665. 10.1257/aer.89.3.649.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–455. 10.1257/jel.48.2.424.

DellaVigna, Stefano, and Elizabeth Linos. 2022. "RCTs to Scale: Comprehensive Evidence From Two Nudge Units." *Econometrica* 90 (1): 81–116. 10.3982/ECTA18709, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18709.

Dove, Kent E. 1999. *Conducting a Successful Capital Campaign: The New, Revised, and Expanded Edition of the Leading Guide to Planning and Implementing a Capital Campaign*. Jossey-Bass, , 2nd edition.

Fox, John A., Jason F. Shogren, Dermot J. Hayes, and James B. Kliebenstein. 1998. "CVM-X: Calibrating Contingent Values with Experimental Auction Markets." *American Journal of Agricultural Economics* 80 (3): 455–465. 10.2307/1244548, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/1244548.

Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29. 10.1111/j.0006-341X.2002.00021.x.

Fuller, Wayne A. 1987. *Measurement Error Models*. Wiley Series in Probability and Mathematical Statistics, New York: Wiley.

Gruber, Jonathan, and Ebonya Washington. 2005. "Subsidies to employee health insurance premiums and the health insurance market." *Journal of Health Economics* 24 (2): 253–276, Publisher: Elsevier.

Hansen, B. B. 2008. "The prognostic analogue of the propensity score." *Biometrika* 95 (2): 481–488. 10.1093/biomet/asn004.

Heckman, James J., and Sergio Urzúa. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156 (1): 27–37. 10.1016/j.jeconom.2009.09.006.

Huck, Steffen, and Imran Rasul. 2011. "Matched fundraising: Evidence from a natural field experiment." *Journal of Public Economics* 95 (5): 351–362. 10.1016/j.jpubeco.2010.10.005.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423. 10.1257/jel.48.2.399.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475. 10.2307/2951620, Publisher: [Wiley, Econometric Society].

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, , 1st edition. 10.1017/CBO9781139025751.

Infosino, William J. 1986. "Forecasting New Product Sales from Likelihood of Purchase Ratings." *Marketing Science* 5 (4): 372–384. 10.1287/mksc.5.4.372, Publisher: INFORMS.

Jackman, Simon. 1999. "Correcting surveys for non-response and measurement error using auxiliary information." *Electoral Studies* 18 (1): 7–27. 10.1016/S0261-3794(98)00039-0.

Jacquemet, Nicolas, Robert-Vincent Joule, Stéphane Luchini, and Jason F. Shogren. 2013. "Preference elicitation under oath." *Journal of Environmental Economics and Management* 65 (1): 110–132. 10.1016/j.jeem.2012.05.004.

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." *Journal of Marketing Research* 26 (3): 336–345. 10.1177/002224378902600307, Publisher: SAGE Publications Inc.

Johannesson, Magnus, Bengt Liljas, and Per-Olov Johansson. 1998. "An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions." *Applied Economics* 30 (5): 643–647. 10.1080/000368498325633, Publisher: Routledge _eprint: https://doi.org/10.1080/000368498325633.

Johansson-Stenman, Olof, and Henrik Svedsäter. 2012. "Self-image and valuation of moral goods: Stated versus actual willingness to pay." *Journal of Economic Behavior & Organization* 84 (3): 879–891. 10.1016/j.jebo.2012.10.006.

Juster, F. Thomas. 1964. *Anticipations and Purchases*. Princeton University Press, , https://press.princeton.edu/books/hardcover/9780691651477/anticipations-and-purchases.

Kang, Min Jeong, Antonio Rangel, Mickael Camus, and Colin F. Camerer. 2011. "Hypothetical and Real Choice Differentially Activate Common Valuation Areas." *Journal of Neuroscience* 31 (2): 461–468. 10.1523/JNEUROSCI.1583-10.2011.

Karlan, Dean, and John A. List. 2007. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment." *American Economic Review* 97 (5): 1774–1793. 10.1257/aer.97.5.1774.

Katz, Jonathan N., and Gabriel Katz. 2010. "Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout." *American Journal of Political Science* 54 (3): 815–835, https://www.jstor.org/stable/27821954, Publisher: [Midwest Political Science Association, Wiley].

Kessler, Judd B., and Alvin E. Roth. 2012. "Organ Allocation Policy and the Decision to Donate." *American Economic Review* 102 (5): 2018–2047. 10.1257/aer.102.5.2018.

Kessler, Judd B., and Alvin E. Roth. 2014. "Getting More Organs for Transplantation." *American Economic Review* 104 (5): 425–430. 10.1257/aer.104.5.425.

Krueger, Alan B., and Ilyana Kuziemko. 2013. "The demand for health insurance among uninsured Americans: Results of a survey experiment and implications for policy." *Journal of Health Economics* 32 (5): 780–793. 10.1016/j.jhealeco.2012.09.005.

Kurz, Mordecai. 1974. "Experimental approach to the determination of the demand for public goods." *Journal of Public Economics* 3 (4): 329–348. 10.1016/0047-2727(74)90004-8.

Lancaster, Kelvin J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy* 74 (2): 132–157, https://www.jstor.org/stable/1828835, Publisher: University of Chicago Press.

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." *Journal of Neuroscience* 31 (1): 118–125. 10.1523/JNEUROSCI.3214-10.2011, Publisher: Society for Neuroscience Section: Articles.

List, John A. 2011. "The Market for Charitable Giving." *Journal of Economic Perspectives* 25 (2): 157–180. 10.1257/jep.25.2.157.

List, John A., and Craig A. Gallet. 2001. "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics* 20 (3): 241–254. 10.1023/A:1012791822804.

List, John A., and Jason F. Shogren. 1998. "Calibration of the difference between actual and hypothetical valuations in a field experiment." *Journal of Economic Behavior & Organization* 37 (2): 193–205. 10.1016/S0167-2681(98)00084-5.

List, John A., and Jason F. Shogren. 2002. "Calibration of Willingness-to-Accept." *Journal of Environmental Economics and Management* 43 (2): 219–233. 10.1006/jeem.2000.1182.

Little, Joseph, and Robert Berrens. 2004. "Explaining Disparities between Actual and Hypothetical Stated Values: Further Investigation Using Meta-Analysis." *Economics Bulletin* 3 (6): 1–13, https://ideas.repec.org/a/ebl/ecbull/eb-03c90005.html, Publisher: AccessEcon.

Loomis, John, Kerri Traynor, and Thomas Brown. 1999. "Trichotomous Choice: A Possible Solution to Dual Response Objectives in Dichotomous Choice Contingent Valuation Questions." *Journal of Agricultural and Resource Economics* 24 (2): 572–583, https://www.jstor.org/stable/40987039, Publisher: Western Agricultural Economics Association.

Louviere, Jordan J. 1993. "Conjoint Analysis." In *Advanced Methods in Marketing Research*, edited by Bagozzi, R. Cambridge: Blackwell Business.

Magnolfi, Lorenzo, Jonathon McClure, and Alan T. Sorensen. 2022. "Triplet Embeddings for Demand Estimation." SSRN Scholarly Paper 4113399, Social Science Research Network, Rochester, NY. 10.2139/ssrn.4113399.

Mansfield, Carol. 1998. "A Consistent Method for Calibrating Contingent Value Survey Data." *Southern Economic Journal* 64 (3): 665–681. 10.2307/1060785, Publisher: Southern Economic Association.

Morrison, Donald G. 1979. "Purchase Intentions and Purchase Behavior." *Journal of Marketing* 43 (2): 65–74. 10.1177/002224297904300207, Publisher: SAGE Publications Inc.

Morwitz, Vicki G., Joel H. Steckel, and Alok Gupta. 2007. "When do purchase intentions predict sales?" *International Journal of Forecasting* 23 (3): 347–364. 10.1016/j.ijforecast.2007.05.015.

Murphy, James J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead. 2005. "A meta-analysis of hypothetical bias in stated preference valuation." *Environmental and Resource Economics* 30 (3): 313–325, Publisher: Springer.

National Oceanic and Atmospheric Association. 1994. "Natural Resource Damage Assessments: Proposed Rules."Technical Report 59, https://www.govinfo.gov/content/pkg/FR-1994-01-07/html/94-225.htm.

Newey, Whitney K., and Daniel McFadden. 1994. "Chapter 36 Large sample estimation and hypothesis testing." In *Handbook of Econometrics*, Volume 4. 2111–2245, Elsevier, . 10.1016/S1573-4412(05)80005-4.

Polak, J., and P. Jones. 1997. "Using Stated-Preference Methods to Examine Traveller Preferences and Responses." *UNDERSTANDING TRAVEL BEHAVIOUR IN AN ERA OF CHANGE*, https://trid.trb.org/view/575079, ISBN: 9780080423906.

Prentice, Ross L. 1989. "Surrogate endpoints in clinical trials: Definition and operational criteria." *Statistics in Medicine* 8 (4): 431–440. 10.1002/sim.4780080407.

Robins, James M., and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90 (429): 122–129. 10.1080/01621459.1995.10476494.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70 (1): 41–55. 10.1093/biomet/70.1.41.

Rothschild, David. 2009. "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* 73 (5): 895–916. 10.1093/poq/nfp082.

Rothschild, David M., and Justin Wolfers. 2011. "Forecasting Elections: Voter Intentions Versus Expectations." *mimeo*, https://www.researchgate.net/publication/256010449_Forecasting_Elections_Voter_Intentions_Versus_Expectations.

Schultz, Henry. 1938. "Theory and measurement of demand." Publisher: The University of Chicago Press.

Shogren, Jason F. 1993. "Experimental Markets and Environmental Policy." *Agricultural and Resource Economics Review* 22 (2): 117–129. 10.1017/S1068280500004706, Publisher: Cambridge University Press.

Shogren, Jason F. 2005. "Chapter 19 Experimental Methods and Valuation." In *Handbook of Environmental Economics*, edited by Mler, Karl-Gran, and Jeffrey R. Vincent Volume 2. of Valuing Environmental Changes 969–1027, Elsevier, . 10.1016/S1574-0099(05)02019-X.

Shogren, Jason F. 2006. "Valuation in the lab." *Environmental and resource Economics* 34 (1): 163–172, Publisher: Springer.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econo-*

metrica 73 (4): 1367–1382. 10.1111/j.1468-0262.2005.00619.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00619.x.

Smith, Alec, B. Douglas Bernheim, Colin F. Camerer, and Antonio Rangel. 2014. "Neural Activity Reveals Preferences without Choices." *American Economic Journal: Microeconomics* 6 (2): 1–36. 10.1257/mic.6.2.1.

Stone, J. R. N. 1954. *The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, 1920-1938*. Volume 1. Cambridge University Press.

Tusche, Anita, Stefan Bode, and John-Dylan Haynes. 2010. "Neural Responses to Unattended Products Predict Later Consumer Choices." *Journal of Neuroscience* 30 (23): 8024–8031. 10.1523/JNEUROSCI.0064-10.2010, Publisher: Society for Neuroscience Section: Articles.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242. 10.1080/01621459.2017.1319839.

Wright, Philip Green. 1928. *The tariff on animal and vegetable oils*. New York: The Macmillan Company, , OCLC: 522698.

Wuthrich, Kaspar, and Ying Zhu. 2021. "Omitted variable bias of Lasso-based inference methods: A finite sample analysis." *arXiv:1903.08704 [econ, math, stat]*, `http://arxiv.org/abs/1903.08704`, arXiv: 1903.08704.

Appendices

A Appendix Figures

Figure A1: Demand Experiment: A typical choice task

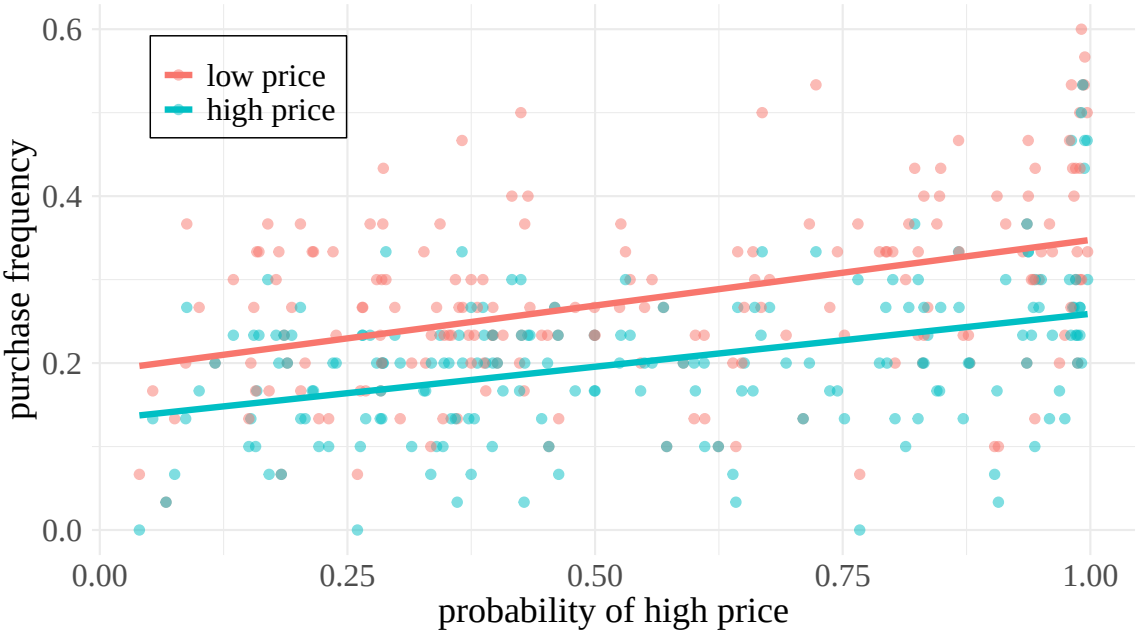


Figure A2: Simulation setup

Potential outcomes corresponding to the high price are in red, and potential outcomes corresponding to the low price are in blue. The curves show the lines of best fit. Snacks likely to be priced at the high price face more demand. This assignment yields the familiar endogeneity problem where the *observed* demand might be higher for high-price snacks than for low-price snack. The probability of high price is determined by our assignment mechanism based on hypothetical WTP. The demand at the low price (red) and high price (blue) is based on the real purchase frequencies in the incentivized experimental group.

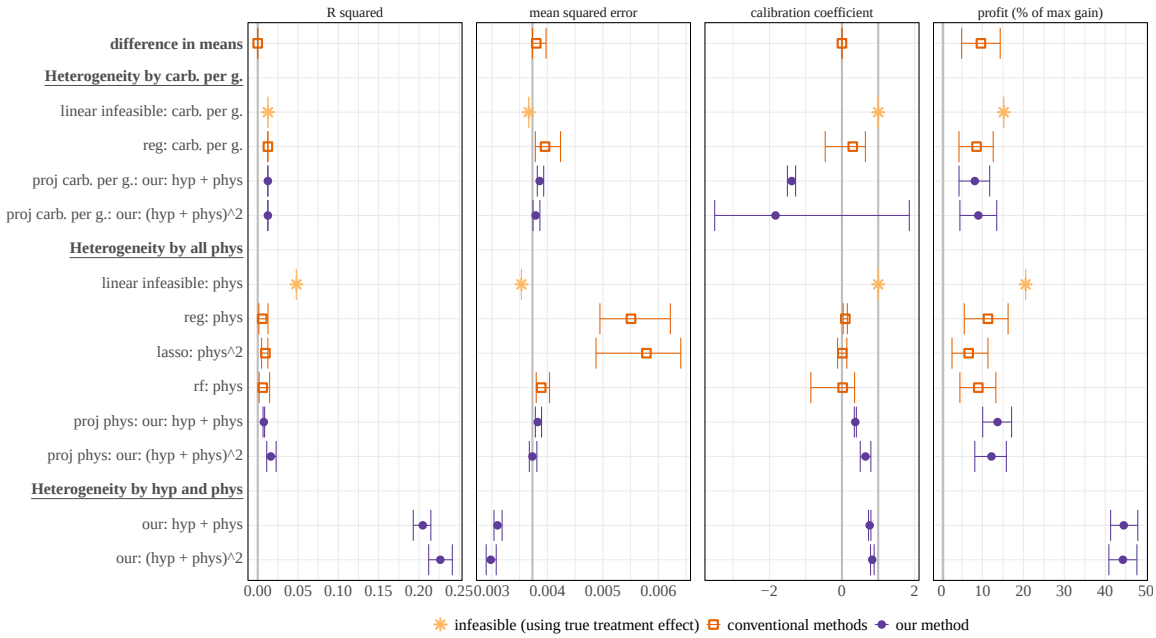


Figure A3: Treatment Effect Heterogeneity: All Estimators

Summary statistics describing how well different estimators describe heterogeneity in treatment effects, with high-dimensional methods. Points show the median statistics across 1,001 simulated samples, and error bars indicate the interquartile range in the simulations.

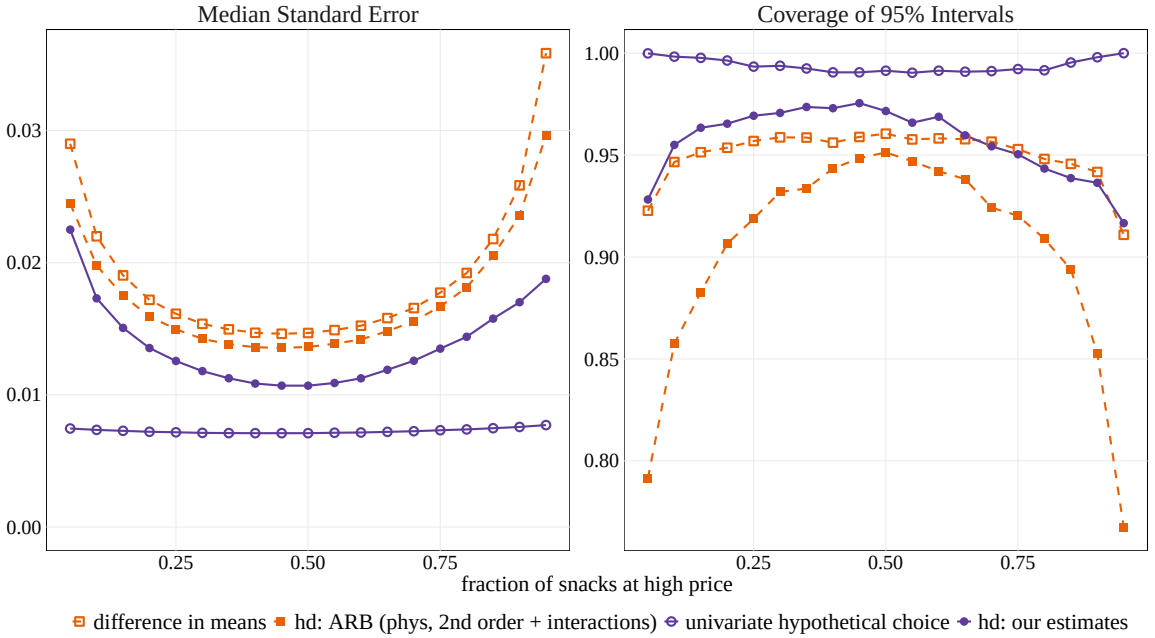
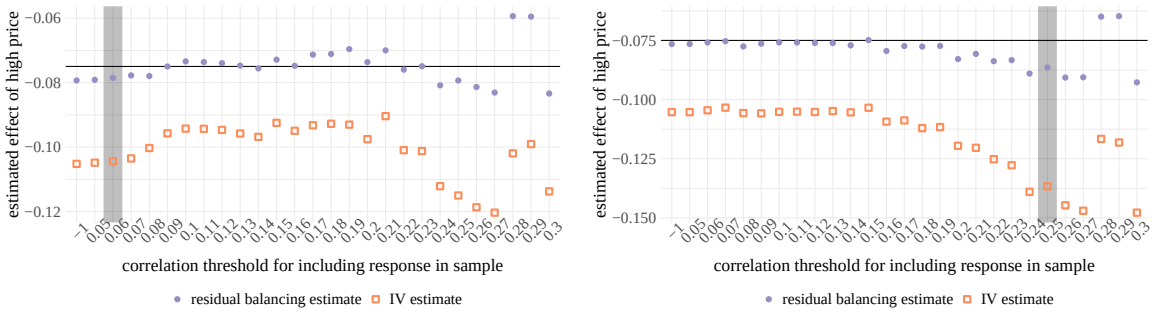


Figure A4: Performance of Estimators by Fraction Treated

Summary statistics describing properties of treatment effect estimators under random assignment. The horizontal axis measures the fraction of snacks observed at the high price. The panels show the median standard error (left) and coverage of nominally 95% confidence intervals (right), across samples differing in treatment assignment.



(a) observing all snacks at high price

(b) observing all snacks at low price

Figure A5: Estimates of the effect of high price by correlation threshold.
The vertical bar indicates the threshold selected by mean square error fit of the Step 1 regression. The horizontal line indicates the true in-sample treatment effect.

B Related Literature

Our approach is related to stated preference (SP) techniques and the contingent valuation method (CVM), which make extensive use of hypothetical choice data (for reviews see Shogren, 2005, 2006; Carson and Hanemann, 2005; Carson, 2012). This literature seeks to predict choices for non-market goods when choice data pertaining to closely related decisions are entirely unavailable (e.g., in the environmental context, to value non-market goods such as pristine coastlines);[58] in contrast, we explore the use of non-choice data as an alternative or supplement to choice data even when the latter are available (but are not ideal).[59]

It is well-established that answers to standard hypothetical questions are systematically biased. Two classes of solutions have been examined. One attempts to "fix" the hypothetical question.[60] Our approach is more closely related to a second class of solutions involving ex post statistical calibration.[61] These techniques exploit statistical relationships between real and hypothetical choices and, like our approach, treat the latter as a predictor rather than a prediction.

The ex post calibration techniques used in the SP/CVM literature differ from ours in several ways. The main distinguishing feature of our approach is that it treats the decision problem as the unit of observation and relates choice distributions to the problem's (subjective) characteristics. In contrast, ex post calibration techniques treat the individual as the unit of observation and relate hypothetical bias to his or her socioeconomic and demographic characteristics. While those techniques account for differences in hypothetical bias across individuals (for a given decision problem), they cannot account for differences across decision problems. Consequently, they are not useful for predicting choice distributions in decision problems that have not yet been observed.[62] On the contrary, List and Shogren

[58] In some cases, the object is to shed light on dimensions of preferences for which real choice data are unavailable by using real and hypothetical choice data in combination; see, e.g., Brownstone et al. (2000) and Small et al. (2005).

[59] Studies that use non-choice data as an alternative and/or supplement to choice data even when the latter are available (but are not ideal) are relatively rare. As an example, consider the problem of estimating the price elasticity of demand for health insurance among the uninsured, who are generally poor and not eligible for insurance through employers. One possibility is to extrapolate from the choices of potentially non-comparable population groups, which also requires one to grapple with the endogeneity of insurance prices, as in Gruber and Washington (2005). Alternatively, Krueger and Kuziemko (2013) attacked the same issue using hypothetical choice data, and reached strikingly different conclusions (i.e., a much larger elasticity).

[60] Methods include the use of (1) certainty scales (as in Champ et al. (1997)), (2) entreaties to behave as if the decisions were real (as in the "cheap-talk" protocol of Cummings and Taylor (1999), or the "solemn oath" protocol of Jacquemet et al. (2013), and (3) "dissonance-minimizing" protocols (as in Blamey et al. (1999), and Loomis et al. (1999), which allow respondents to express support for a public good while also indicating a low WTP).

[61] See Kurz (1974), Shogren (1993), Blackburn et al. (1994), National Oceanic and Atmospheric Association (1994), Fox et al. (1998), List and Shogren (1998, 2002), and Mansfield (1998).

[62] Indeed, unlike our analysis, existing ex post calibration studies do not generally focus on out-of-sample predictive performance. Nor do they run the types of "horse races" between choice-based and non-choice-based prediction methods that reveal whether these methods have merit in settings where (imperfect) choice data are

(1998; 2002) emphasize that hypothetical bias is context-specific, so that individual-level calibration does not reliably transfer from one setting to another.[63] Yet psychological studies also suggest that hypothetical bias is systematically related to measurable factors that vary across decision problems (e.g., Ajzen et al. (2004), and Johansson-Stenman and Svedsäter (2012)). Our approach allows us to adjust for factors affecting the degree of hypothetical bias that vary across decision problems by including other appropriate non-choice responses, such as questions that elicit norms or image concerns.

An additional advantage of conducting our analysis at the level of the decision problem is that we can assess non-choice responses using different groups of subjects. In contrast, in ex post calibration studies, subjects make real choices after making hypothetical ones, which introduces the possibility of cross-contamination. Our ability to obtain independent non-choice responses with distinct groups also allows us to employ, in a single specification, combinations of predictors that include multiple versions of hypothetical choices (e.g., standard, certainty scaled, and cheap-talk variants) along with other subjective ratings, and to determine whether those measures have independent and complementary predictive power. In contrast, the aforementioned studies calibrate hypothetical choices one version at a time.

A separate pertinent strand of research within the SP/CVM literature involves meta-analyses (Carson and Hanemann, 2005; List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005). Unlike the ex post calibration literature, those studies attempt to find variables that account for the considerable variation in hypothetical bias across contexts and goods. However, they are primarily concerned with evaluating the effects of diverse experimental methods on hypothetical bias,[64] rather than with assessing out-of-sample predictive accuracy, as we do.

Stepping away from SP data, portions of the neuroeconomics literature seek to predict choices from neural and/or physiological responses. Smith et al. (2014) focus specifically on passive non-choice neural reactions, and provide proof-of-concept that those types of reactions predict choices.[65] Separately, in the literature on subjective well-being, two papers explore the relationships between forward-looking statements concerning happiness and/or satisfaction and hypothetical choices (Benjamin et al., 2012, 2014), which motivates our use of such variables to predict real choices.

Turning to other disciplines, the marketing literature has examined stated intentions as predictors of purchases (see, e.g., Juster, 1964; Morrison, 1979; Infosino, 1986; Jamieson

also available.

[63] Blackburn et al. (1994) provide somewhat mixed evidence on portability, but their analysis is limited to two goods.

[64] One exception is that they point to a systematic difference in hypothetical bias for public and private goods.

[65] See also Tusche et al. (2010) and Levy et al. (2011).

and Bass, 1989). Its relationship to our work is similar to that of the SP/CVM literature on ex post calibration techniques in that the object, once again, is to derive individual-specific predictions for a given good, with cross-good differences addressed through meta-analysis (e.g., Morwitz et al., 2007). Marketing scholars also routinely use SP data (derived from "choice experiments" involving hypothetical choices over multiple alternatives) to estimate preference parameters in the context of a single choice problem (see Louviere, 1993; Polak and Jones, 1997; Ben-Akiva et al., 1994; Alpizar Rodriguez et al., 2003, for useful reviews). Our analysis provides methods for potentially improving those data inputs. There are also parallels to our work in the political science literature, particularly concerning the prediction of voter turnout and election results, e.g., from surveys and polls (as in Jackman (1999), and Katz and Katz (2010)). As in our approach, the object is to predict aggregate outcomes rather than individuals' choices, and a range of potential predictors (in addition to hypothetical choices or intentions) are sometimes considered. For example, Rothschild and Wolfers (2011) find that questions concerning likely electoral outcomes (i.e., how others will vote) are better predictors than stated intentions.[66] The problem is substantively different, however, in that surveys and polls ask voters about real decisions that many have made, plan to make, or are in the process of making, instead of measuring non-choice reactions to choice problems that respondents view as hypothetical.

C An explicit model of underlying processes

In this section, we provide an explicit model of underlying processes and clarify the nature of our statistical assumptions within that context. It is worth emphasizing that we intend this model only as an illustration of the types of processes for which our assumptions might hold.

C.1 Treatments and choices

We consider applications with settings (indexed $j = 1, \ldots, J$, representing treatment units such as goods, geographical jurisdictions, or markets) in which a set of individuals (indexed i) make choices, Y_{ij}, subject to the treatment assigned to that setting, $W_j \in \mathbb{W}$. The set of individuals may be identical across settings, overlapping between settings, or disjoint.[67]

The treatment assigned to setting j depends on its stable characteristics $\boldsymbol{X}_j$ and $\boldsymbol{\eta}_j$, which are respectively observable and unobservable to the econometrician, and typical conditions $\boldsymbol{\xi}_{ij}^{typ} \sim F_j^{typ}$ that may vary across individuals. Thus, $W_j = W_j(\boldsymbol{X}_j, \boldsymbol{\eta}_j, F_j^{typ})$.

[66] Some studies also use prediction markets (e.g., Rothschild, 2009), which (in effect) elicit investors' incentivized forecasts of electoral outcomes.

[67] If we take the the set of individuals as given (i.e., condition on them) and consider randomness only from treatment assignment and the realization of actual choices (as discussed below), identical or overlapping sets of individuals do not necessarily introduce statistical dependence across settings.

Individual i's choice in setting j depends on the treatment, stable characteristics of the setting, $\boldsymbol{X}_j$ and $\boldsymbol{\eta}_j$, and unobserved *realized* conditions $\boldsymbol{\xi}_{ij} \sim F_j$ that i experiences in setting j. Thus, $Y_{ij} = Y(W_j, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$.[68] We are primarily concerned with either binary choices $Y_{ij} \in \{0, 1\}$ or continuous choices $Y_{ij} \in \mathbb{R}$.

Endogeneity may arise from two sources. First, unobservable factors $\boldsymbol{\eta}_j$ affect both treatment and choices. Second, some components of the draws $\boldsymbol{\xi}_{ij}^{typ}$ may be unobserved, and there is a relationship between the distribution F_j^{typ} that affects treatment and the distribution F_j that affects choices.

The average outcome in setting j with treatment state w is

$$Y_j^{typ}(w) = \int Y(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij}^{typ}) \, \mathrm{d}F_j^{typ}$$

under typical conditions, and is

$$Y_j(w) = \int Y(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij}) \, \mathrm{d}F_j = Y_j^{typ}(w) + \epsilon_j(w)$$

under realized conditions, where the error term $\epsilon_j(w)$ reflects the difference between distributions F_j and F_j^{typ}. Since treatment assignment is based on choices under typical conditions, it is natural to assume that this error is orthogonal to treatment, given the determinants of treatment,

$$W_j \perp\!\!\!\perp \{\epsilon_j(w)\}_{w \in \mathbb{W}} \mid \left\{ Y_j^{typ}(w) \right\}_{w \in \mathbb{W}}.$$

C.2 Motivations

We conceptualize choice as resulting from the psychological *motivations,* $\boldsymbol{\theta}_{ij}(w)$, that arise for individual i in setting j under treatment state w:

$$Y_{ij}(w) = Y^*(\boldsymbol{\theta}_{ij}(w))$$

We assume that these motivations reflect the treatment as well as the observed and unobserved characteristics of the individual and the setting: $\boldsymbol{\theta}_{ij}(w) = \boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$ or $\boldsymbol{\theta}_{ij}(w) = \boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij}^{typ})$, depending on whether the motivations are formed under actual or under typical conditions. At this level of generality, external conditions, including the

[68] If the actor choosing the treatment can envision and account for variation in the potential realizations of F_j, then in principle one should define F_j^{typ} to account for that variation, rather than limiting it to the distribution arising in a typical condition. To accommodate that alternative assumption, one would have to elicit a distribution of responses for each individual rather than a typical response, which would likely prove challenging. We therefore proceed under the assumption that the distribution of responses under typical conditions captures the information relevant to treatment selection, and that the variability of the realized distribution is of second-order importance with respect to selection.

treatment, affect choices only indirectly through motivations. This exclusion restriction should not be controversial, inasmuch as choices are governed by internal representations of decision problems. It follows that

$$Y_j^{typ}(w) = \int Y^*(\boldsymbol{\theta}_{ij}(w)) \, \mathrm{d}F_j^{typ,\boldsymbol{\theta}(w)},$$

where $F_j^{typ,\boldsymbol{\theta}(w)}$ is the marginal distribution of $\boldsymbol{\theta}_{ij}(w)$ for setting j and treatment status w implied by the distribution of $\boldsymbol{\xi}_j^{typ}$ under typical conditions, F_j^{typ}.

For the sake of simplicity, we focus here on the case of binary treatments, $W_j \in \{0, 1\}$, and assume we can write the integral in the preceding equation as a stable linear function of variables $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(w)$ describing features of the marginal distribution $F_j^{typ,\boldsymbol{\theta}(w)}$, such as moments and percentiles. For now, we also assume $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(w)$ is perfectly observable for all settings and treatment states.

Assume for the moment that we observe the potential outcomes $Y_j^{typ}(w)$ under typical conditions in *both* treatment states. Suppose we regress $Y_j^{typ}(w)$ on the distributional characteristics $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(w)$, pooling observations from all settings and treatment conditions, and then use the estimated equations to compute fitted choices, $\hat{Y}_j(0)$ and $\hat{Y}_j(1)$. As long as we select a functional specification with sufficient flexibility to accommodate the variation in conditional expectations, the treatment effect under typical conditions, $Y_j^{typ}(1) - Y_j^{typ}(0)$, will equal the fitted treatment effect, $\hat{Y}_j(1) - \hat{Y}_j(0)$.[69]

In practice, instead of $Y_j^{typ}(0)$ and $Y_j^{typ}(1)$, we observe $Y_j(W_j)$, the outcome for setting j, under realized rather than typical conditions, and only for the treatment condition that actually prevails. We can nevertheless employ our proposed method: that is, we can run the same regression using the available data (i.e., regress $Y_j(W_j)$ on $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$), use it to construct a fitted value and a prediction, $\hat{Y}_j(1)$ and $\hat{Y}_j(0)$, and then compute $\hat{Y}_j(1) - \hat{Y}_j(0)$ exactly as before. If the distributions of the covariates $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(1 - W_j)$ have sufficient overlap, we can proceed nonparametrically; otherwise, extrapolation requires a correct functional form.

When we observe data only for the actual treatment states, those observations are systematically selected. However, by assumption, the treatment depends only on the features of the setting and typical conditions $(\boldsymbol{X}_j, \boldsymbol{\eta}_j, F_j^{typ})$. Because these factors affect outcomes only through $\boldsymbol{\theta}_{ij}(W_j)$, which we have assumed is observed, the treatment is unconfounded. It follows that observing just one of the potential outcomes for each setting does not cause systematic biases. Formally, the covariates $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(0)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(1)$ are *balancing scores* in

[69] With multi-valued treatments, one could similarly fit the choices $Y_j(w)$ for all relevant treatment states $w \in \mathbb{W}$, and aggregate these predictions into a meaningful statistic such as an average derivative or elasticity.

the sense of Rosenbaum and Rubin (1983).[70]

The other difference between our procedure and the (infeasible) fitted treatment effect procedure is that we use data on $Y_j(W_j)$ rather than $Y_j^{typ}(W_j)$. However, we will still correctly estimate the relationship between $Y_j^{typ}(W_j)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$ as long as the differences between (average) outcomes under realized and typical conditions, $\epsilon_j(W_j)$, are not systematically related to the distributions of typical intentions $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$. This assumption is plausible if the difference reflects sampling, or if conditions modulate baseline intentions (and hence outcomes) in a similar way across settings. It is particularly natural for cases involving linear relationships between choices and measured intentions: if $\epsilon_j(W_j)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$ were correlated, then presumably F_j^{typ} would not reflect the most representative conditions.

It follows that the differences between the our procedure and the fitted treatment effect procedure are innocuous under reasonable assumptions. The requirements of the method therefore largely boil down to whether it is possible to measure motivations sufficiently well.

While motivations are necessarily measured imperfectly, that is not necessarily problematic. Typically, we elicit motivations based on answers to hypothetical questions, $\boldsymbol{H}_{kj}(w)$, from some set of individuals similar to but distinct from those who make actual choices (indexed k). As discussed in the main text, we use a distinct sample to avoid real choices contaminating hypothetical evaluations, or vice versa. We regress $Y_j(W_j)$ on $\boldsymbol{D}_j^{typ,\boldsymbol{H}}(W_j)$ rather than $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(W_j)$; the procedure is otherwise the same. The validity of this approach depends on how hypothetical motivations for survey respondents relate to typical motivations for decision makers.

D Additional estimation results

D.1 Proof of Theorem 1

The data are a random sample of independent observations $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j)_{j=1}^J$ where $Y_j \in \mathbb{R}$, $W_j \in \{0,1\}$, and $\boldsymbol{H}_j(1), \boldsymbol{H}_j(0) \in \mathbb{R}^{Q_H}$ as well as $\boldsymbol{X}_j \in \mathbb{R}^{Q_X}$ are row vectors. For ease of notation, we define row vectors $\boldsymbol{Z}_j(w) = [\boldsymbol{H}_j(w),\ \boldsymbol{X}_j] \in \mathbb{R}^Q$ with $Q = Q_H + Q_X$. Let $\boldsymbol{Z}_j = \boldsymbol{Z}_j(W_j)$. The estimator proceeds in two steps: first, regress outcomes Y_j on hypothetical evaluations and fixed characteristics $\boldsymbol{Z}_j$. Second, take the estimated coefficients on $\boldsymbol{Z}_j$, say $\hat{\boldsymbol{\delta}} = [\hat{\boldsymbol{\beta}}^T,\ \hat{\boldsymbol{\gamma}}^T]^T$, and calculate $\hat{\tau} = \frac{1}{J}\sum_{j=1}^J (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\hat{\boldsymbol{\beta}}$.

[70] See also recent work on the prognostic score (Hansen, 2008).

Write the two-step estimator in a single GMM framework with moments

$$g(y, \boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}, \tau, \boldsymbol{\delta}) = \tau - (\boldsymbol{z}_1 - \boldsymbol{z}_0)\boldsymbol{\delta}$$
$$\boldsymbol{m}(y, \boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}, \tau, \boldsymbol{\delta}) = \boldsymbol{z}'(y - \boldsymbol{z}\boldsymbol{\delta})$$

By Assumptions 1, 2, and 3,

$$\mathbb{E}(g(Y_j, \boldsymbol{Z}_j(0), \boldsymbol{Z}_j(1), \boldsymbol{Z}_j, \tau^*, \boldsymbol{\delta}^*) = 0$$

where $\tau^* = \mathbb{E}(Y_j(1) - Y_j(0))$ and $\boldsymbol{\delta}^* \equiv [\boldsymbol{\beta}_{0,0}^T,\ \boldsymbol{\gamma}_0^T]^T = [\boldsymbol{\beta}_{1,1}^T,\ \boldsymbol{\gamma}_1^T]^T$ with $\boldsymbol{\beta}_{w,w}$ and $\boldsymbol{\gamma}_w$ as specified in Assumption 3. The equality holds by Assumptions 1 and 2. Assumption 1 further implies that $\boldsymbol{\beta}_{0,1} = \boldsymbol{\beta}_{1,0} = 0$. By Assumptions 1, 2, 3 and 4 and random sampling,

$$\mathbb{E}(\boldsymbol{m}(Y_j, \boldsymbol{Z}_j(0), \boldsymbol{Z}_j(1), \boldsymbol{Z}_j, \tau^*, \boldsymbol{\delta}^*) = \mathbf{0}_{Q\times 1}$$

where $\mathbf{0}_{Q\times 1}$ is the $Q \times 1$ zero matrix.

Let $\boldsymbol{\psi} = (g^T, \boldsymbol{m}^T)^T$ be the vector stacking these moments. Then $\mathbb{E}(\boldsymbol{\psi}) = 0$.

Define

$$\begin{aligned}\boldsymbol{\Gamma} &= \mathbb{E}\Big(\frac{\partial \psi(Y_j, \boldsymbol{Z}_j(0), \boldsymbol{Z}_j(1), \boldsymbol{Z}_j, \tau^*, \boldsymbol{\delta}^*)}{\partial(\tau, \boldsymbol{\delta}^T)}\Big) \\ &= \mathbb{E}\left(\begin{bmatrix} 1 & -(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)) \\ \mathbf{0}_{Q\times 1} & -\boldsymbol{Z}_j'\boldsymbol{Z}_j \end{bmatrix}\right)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Psi} &= \mathbb{E}(\boldsymbol{\psi}\boldsymbol{\psi}') = \mathbb{E}\left(\begin{bmatrix} g^2 & g\boldsymbol{m}^T \\ g\boldsymbol{m} & \boldsymbol{m}\boldsymbol{m}^T \end{bmatrix}\right) \\ &= \mathbb{E}\left(\begin{bmatrix} (\tau^* - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta}^*)^2 & \boldsymbol{Z}_j(\tau^* - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta}^*)(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*) \\ \boldsymbol{Z}_j^T(\tau^* - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta}^*)(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*) & \boldsymbol{Z}_j^T\boldsymbol{Z}_j(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*)^2 \end{bmatrix}\right)\end{aligned}$$

Then, under standard regularity conditions, the asymptotic distribution of $(\hat{\tau}, \hat{\boldsymbol{\delta}})$ is

$$\sqrt{J}\left(\begin{bmatrix}\hat{\tau} \\ \hat{\boldsymbol{\delta}}\end{bmatrix} - \begin{bmatrix}\tau^* \\ \boldsymbol{\delta}^*\end{bmatrix}\right) \to^d N\Big(\mathbf{0}_{(1+Q)\times 1},\ \boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}(\boldsymbol{\Gamma}^T)^{-1}\Big)$$

The asymptotic variance of $\hat{\tau}$ is given by the $(1,1)$ element of the variance matrix $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}(\boldsymbol{\Gamma}')^{-1}$. By Newey and McFadden (1994, Theorem 6.1),

$$\sqrt{J}(\hat{\tau} - \tau) \to^d N(0, V_\tau)$$

where

$$V_\tau = \mathbb{E}(g^2) + \mathbb{E}\Big(\frac{\partial g}{\partial \boldsymbol{\delta}}\Big)^T \boldsymbol{V}^{\text{ols}} \mathbb{E}\Big(\frac{\partial g}{\partial \boldsymbol{\delta}}\Big) - 2\mathbb{E}\Big(\frac{\partial g}{\partial \boldsymbol{\delta}}\Big)^T \Big(\mathbb{E}\Big(\frac{\partial \boldsymbol{m}}{\partial \boldsymbol{\delta}^T}\Big)^{-1}\Big)\mathbb{E}(g\boldsymbol{m})$$

with $\boldsymbol{V}^{\text{ols}} = \mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j(y-\boldsymbol{Z}_j\boldsymbol{\delta}^*)^2\Big)\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}$ the $Q \times Q$ asymptotic variance matrix of $\hat{\boldsymbol{\delta}}$ in the first-step OLS regression. Substituting the moment functions g and $\boldsymbol{m}$ and their derivatives, obtain

$$\begin{aligned} V_\tau = \mathbb{E}\Big((\tau^* - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta}^*)^2\Big) \\ + \mathbb{E}\Big(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)\Big)\boldsymbol{V}^{\text{ols}}\mathbb{E}\Big(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)\Big)^T \\ - 2\mathbb{E}\Big(\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0)\Big)\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}_j^T(\tau^* - (\boldsymbol{Z}_j(1) - \boldsymbol{Z}_j(0))\boldsymbol{\delta}^*)(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*)\Big) \end{aligned}$$

D.2 Estimators for high-dimensional evaluations and non-linear relationships

We develop a machine learning estimator for cases involving linearity in high-dimensional hypothetical evaluations.

Let $\boldsymbol{Z}_j(w) = g(\boldsymbol{H}_j(w), \boldsymbol{X}_j)$ be the covariate vector for setting j, including predictors $\boldsymbol{H}_j(w)$ for treatment state $w \in \{0, 1\}$ and fixed characteristics $\boldsymbol{X}_j$, as well as any transformations, higher order terms, and interactions. Analogously to a Taylor expansion, a linear combination of a sufficiently large number of transformations can approximate complicated nonlinear functions.

Although LASSO is a popular estimator for applied work, LASSO coefficient estimates can suffer from biases due to under-selection in finite samples (for instance, Wuthrich and Zhu, 2021). We propose a high-dimensional counterpart involving a variant of approximate residual balancing (ARB, Athey et al., 2018), which removes such biases for aggregate predictions.

Computation of the estimator $\hat{\tau}_{\text{arb}}$ involves the following steps:

Step 1a. Using LASSO, estimate the relationship between the realized outcome Y_j and the covariates $\boldsymbol{Z}_j = \boldsymbol{Z}_j(W_j)$ for the realized treatment state:

$$\hat{\boldsymbol{\delta}}_{\text{lasso}} = \arg\min_{\delta} \sum_{j=1}^{J} \Big(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}\Big)^2 + \lambda\|\boldsymbol{\delta}\|_1$$

where the tuning parameter λ is chosen through cross-validation.

Step 1b. Compute approximate balancing weights

$$\boldsymbol{\rho}^t = \arg\min_{\tilde{\boldsymbol{\rho}}\in\mathbb{R}^N} \zeta\|\tilde{\boldsymbol{\rho}}\|_2^2 + (1-\zeta)\|\overline{\boldsymbol{Z}(1)} - \boldsymbol{Z}^T\tilde{\boldsymbol{\rho}}\|_\infty^2$$
$$\text{subject to: } \sum_{j=1}^J \tilde{\rho}_j = 1; \quad \forall j:\ 0 \le \tilde{\rho}_j \le J^{-2/3}$$
$$\boldsymbol{\rho}^c = \arg\min_{\tilde{\boldsymbol{\rho}}\in\mathbb{R}^N} \zeta\|\tilde{\boldsymbol{\rho}}\|_2^2 + (1-\zeta)\|\overline{\boldsymbol{Z}(0)} - \boldsymbol{Z}^T\tilde{\boldsymbol{\rho}}\|_\infty^2$$
$$\text{subject to: } \sum_{j=1}^J \tilde{\rho}_j = 1; \quad \forall j:\ 0 \le \tilde{\rho}_j \le J^{-2/3}$$

where $\boldsymbol{Z}$ stacks the covariates $\boldsymbol{Z}_j$ for all decision problems, and $\overline{\boldsymbol{Z}(w)} = \frac{1}{J}\sum_{j=1}^J \boldsymbol{Z}_j(w)$ for $w \in \{0,1\}$. Athey et al. (2018) sets the tuning parameter $\zeta = 0.5$ as a default.

Step 2. Estimate the average treatment effect as

$$\hat{\tau}_{\text{arb}} = \Big(\overline{\boldsymbol{Z}(1)} - \overline{\boldsymbol{Z}(0)}\Big)\hat{\boldsymbol{\delta}}_{\text{lasso}} + \sum_{j=1}^J (\boldsymbol{\rho}_j^t - \boldsymbol{\rho}_j^c)\Big(Y_j - \boldsymbol{Z}_j\hat{\boldsymbol{\delta}}_{\text{lasso}}\Big)$$

If we included only the first term in Step 2, the procedure would be analogous to replacing OLS with LASSO in our low-dimensional procedure. The second term in Step 2 addresses the biases associated with high-dimensional estimation and penalization by adding weighted prediction errors from Step 1a. The particular weights $\boldsymbol{\rho}^t$ and $\boldsymbol{\rho}^c$, computed in Step 1b, are meant to reduce estimation errors for $\mathbb{E}\big(\mathbb{E}(Y_j(1)|\boldsymbol{Z}_j(1))\big)$ and $\mathbb{E}\big(\mathbb{E}(Y_j(0)|\boldsymbol{Z}_j(0))\big)$ in the first term of Step 2, under the assumption of linearity.[71]

Restricting to respondents who are most skilled When we use ARB in the setup that identifies the most informative respondents in Section 4.4, we augment the procedure as follows. In Step 1a, rather than using a single penalized (LASSO) regression;for a given threshold, we split the responses of respondents in half, and aggregate responses for each observation within each half. We then estimate separate instrumental variables regressions for each threshold, using Y_j as the outcome variable and instrumenting for the aggregate hypothetical evaluations of one half with the aggregate hypothetical evaluations of the other half. We then reverse the roles of the first and second halves and average the estimated coefficients from the two regressions. For a given threshold r, this creates an estimated

[71] Specifically, the objective functions in Step 1b have two parts. Introducing $\|\tilde{\boldsymbol{\theta}}\|_2^2$ reduces the variance of the estimator by penalizing deviations from equal weights. Introducing $\|\overline{\boldsymbol{Z}(w)} - \boldsymbol{Z}^T\tilde{\boldsymbol{\theta}}\|_\infty^2$ limits bias under the assumption of linearity by penalizing the deviations from exact covariate balance between the weighted covariates $\boldsymbol{Z}_j$ used in estimation in Step 1 and the average covariates $\overline{\boldsymbol{Z}(w)}$ used to predict outcomes in the first part of Step 2; this term is the maximum (across covariates) squared deviation between these average covariates.

coefficient vector, say $\hat{\boldsymbol{\delta}}^r$. When selecting one of the thresholds r, Step 1a resembles "subset selection" as an alternative to the LASSO regression in the original version of ARB.

We take, for the purpose of Step 1b and Step 2 the vectors $\boldsymbol{Z}_j(w)$ and $\boldsymbol{Z}_j$ to be the collection of average covariates of *all* thresholds. The estimator given choice of threshold r^*, is then $\hat{\tau}^{r^*}_{\text{arb}} = \left(\overline{\boldsymbol{Z}(1)} - \overline{\boldsymbol{Z}(0)}\right)\hat{\boldsymbol{\delta}}^{r^*} + \sum_{j=1}^{J}(\rho_j^t - \rho_j^c)\left(Y_j - \boldsymbol{Z}_j\hat{\boldsymbol{\delta}}^{r^*}\right)$. The second term ensures that the estimate is close to the true effect even if the threshold r^* is not selected correctly in finite samples, as long as the true model is linear the average hypothetical evaluations under the different thresholds. Appendix F.2 describes a way to choose a threshold; however, in practice, we find that this estimator reduces the importance of selecting between thresholds, because these robust estimates are similar across all choices of r.

D.2.1 Theoretical Results

The formal analysis of $\hat{\tau}_{\text{arb}}$ requires an additional overlap assumption. Overlap is commonly assumed for non-parametric estimators in causal inference, but in our setting a noticeably weaker version, which we term *evaluations overlap*, suffices:

Assumption 5. Evaluations overlap. *For each value of the predictors, pooling treatment states, the probability of treatment is bounded away from 0 and 1. Specifically, if $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are the supports of the distributions of predictors in the control and treatment states, respectively, then for all $\boldsymbol{z} \in (\mathbb{Z}_0 \cup \mathbb{Z}_1)$, we have for some $\eta > 0$ at least one of*

$$\Pr(W_j = 1 \mid \boldsymbol{Z}_j(0) = \boldsymbol{z}) < 1 - \eta$$

or

$$\eta < \Pr(W_j = 1 \mid \boldsymbol{Z}_j(1) = \boldsymbol{z})$$

A sufficient condition for this assumption is that, for any value of the predictors $\boldsymbol{z} \in (\mathbb{Z}_0 \cup \mathbb{Z}_1)$, we observe (a growing number of) settings j for which the hypothetical evaluations corresponding to the realized treatment state coincide with $\boldsymbol{z}$, i.e., $\boldsymbol{Z}_j(W_j) = \boldsymbol{z}$. The overlap assumption is therefore substantially weaker than for standard treatment effects estimators. In particular, Assumption 5 can hold even when there is no variation in treatment assignment. Notably, in that special case, unconfoundedness (Assumption 4) is also satisfied trivially.

Under the preceding assumptions and regularity conditions, the following theorem demonstrates that our estimator $\hat{\tau}_{\text{arb}}$ is consistent for the average treatment effect, and asymptotically normal with straightforward standard errors.

Theorem 2. *Suppose our Assumptions 1, 2, 3 (here linearity in high-dimensional covariates $\boldsymbol{Z}_j(w)$ rather than $\boldsymbol{H}_j(w)$), 4, and 5, as well as assumptions from Athey et al. (2018) – exact*

sparsity Assumption 4, regularity conditions on the covariates $\boldsymbol{Z}$ of Assumption 7, regularity conditions on the (potentially heteroskedastic) regression noise in Corollary 2 – hold. Suppose further that we use the estimator $\hat{\tau}_{\text{arb}}$ with a hard constraint replacing the Lagrange form penalty on the imbalance in our Step 1b (analogous to the constraint in Theorem 2 of Athey et al. (2018)). Then the estimator $\hat{\tau}_{\text{arb}}$ is asymptotically normal with

$$\frac{\hat{\tau}_{\text{arb}} - \tau}{\sqrt{\hat{V}_{\text{arb}}}} \to \mathcal{N}\Big(0, 1\Big)$$

where $\hat{V}_{\text{arb}} = \sum_{j=1}^{N} (\rho_j^t - \rho_j^c)^2 (Y_j - \boldsymbol{Z}_j \hat{\boldsymbol{\delta}}_{\text{lasso}})^2$.[72]

Proof: The result follows from Lemma 2 of Athey et al. (2018) by noting that our unconfoundedness Assumption 4 has the same implication as their Assumption 1 for this estimation step, our Assumptions 1, 2, and 3 jointly imply their Assumption 2, and our overlap Assumption 5 is identical to their Assumption 6 after rewriting our variables according to their setup. Their condition on the limit of the odds ratio is not needed in our setting because we observe covariates $\boldsymbol{Z}_j(0)$ and $\boldsymbol{Z}_j(1)$ and an outcome Y_j for all decision problems irrespective of treatment assignment. The two weights $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^c$ separately balance for estimation of the mean of treated and the mean of control potential outcomes, as in the "Proof of Lemma 9" in their on-line appendix for the mean of the control, and the difference $\boldsymbol{\theta}^t - \boldsymbol{\theta}^c$ takes the role of $\boldsymbol{\theta}$ in the "Proof of Corollary 6" in their on-line appendix..

D.3 Nonparametric identification

While our main estimators make assumptions about functional form, such assumptions are not necessary to identify treatment effects:

Theorem 3. *The average effect of the treatment, $\tau = \mathbb{E}(Y_j(1) - Y_j(0))$, is nonparametrically identified under Assumptions 1, 2, 4, and 5.*

Proof: $\mathbb{E}(Y_j(1) - Y_j(0)) = \mathbb{E}\big(\mathbb{E}(Y_j(1) - Y_j(0) \mid \boldsymbol{H}_j(1), \boldsymbol{H}_j(0), \boldsymbol{X}_j)\big)$ by the law of iterated expectations. The next steps hold for $w \in \{0, 1\}$. By Assumption 1, $\mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(1), \boldsymbol{H}_j(0), \boldsymbol{X}_j) = \mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(w), \boldsymbol{X}_j)$. $\mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) = \mathbb{E}(Y_j \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}, W_j = w)$ by unconfoundedness Assumption 4. $\mathbb{E}(Y_j \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}, W_j = w) = \mathbb{E}(Y_j \mid \boldsymbol{H}_j(W_j) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$ is identified by Assumptions 2 and 5 for all relevant levels of $\boldsymbol{h}$ and $\boldsymbol{x}$.

[72] In contrast to the variance in Theorem 1, the variance estimator $\hat{V}_{\text{arb}}$ in Theorem 2 is conditional on hypothetical evaluations. Specifically, for a fixed sample size, the weights $(\rho_j^t - \rho_j^c)$ are deterministic (fixed) under sampling of outcomes Y_j conditional on covariates $\boldsymbol{Z}_j$ and treatment assignment W_j. Hence, if one is specifically interested in comparing the estimated standard errors across our low-dimensional and high-dimensional methods, the proper counterpart to $\hat{V}_{\text{arb}}$ from Theorem 2 is the second term of $\hat{V}_{\text{p}}$ from Theorem 1.

Theorem 3 says that we can estimate treatment effects without making functional form assumptions. We therefore view parametric assumptions, such as linearity, primarily as useful approximations, but our approach is not fundamentally tied to them.

D.4 Doubly robust estimators

For an alternative doubly robust estimator along the lines of Robins and Rotnitzky (1995) and Chernozhukov et al. (2018) using our Assumptions 1, 2, and 4 it is easy to verify that the following moment condition satisfies the Neyman orthogonality condition:

$$\psi(y, w, \boldsymbol{h}_1, \boldsymbol{h}_0, \boldsymbol{x}) = \mu(\boldsymbol{h}_1, \boldsymbol{x}) - \mu(\boldsymbol{h}_0, \boldsymbol{x}) + \frac{w}{e_1(\boldsymbol{h}_1, \boldsymbol{x})}\Big(y - \mu(\boldsymbol{h}_1, \boldsymbol{x})\Big) - \frac{1-w}{e_0(\boldsymbol{h}_0, \boldsymbol{x})}\Big(y - \mu(\boldsymbol{h}_0, \boldsymbol{x})\Big)$$

where μ is the relationship between outcome and hypothetical evaluations of the realized treatment state, and $e_w(\boldsymbol{h}, \boldsymbol{x}) = \Pr(W_j = w | \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$ for $w \in \{0, 1\}$ is the probability that decision problem j is observed in state w conditional on the hypothetical evaluations of that state and fixed characteristics. To avoid biases, μ and e_w should be estimated using cross-fitting. Under suitable conditions for the machine learning estimators of choice for μ and e_w, such a doubly robust estimator may perform well. Note, however, that our framework does not suggest that we are well-positioned to correctly specify a propensity score conditional on hypothetical evaluations. Although this doubly robust moment uses the same structural Assumptions 1 and 2, it also requires a standard overlap assumption bounding conditional treatment probabilities away from 0 and 1. Consequently, it cannot be used to estimate the effect of a treatment that has not been implemented. It is an interesting question whether it is possible to construct a doubly robust estimator of this type that retains the advantages of our parametric and residual balancing estimators.

E Snack Demand Application

E.1 Groups

Group R (30 subjects): Subjects made real choices using the strategy method. Each item appeared twice, once with a price of 25 cents and once with a price of 75 cents. In each case, the subject had to decide whether to buy the item at the specified price. The subject was told that, prior to stage 2 of the experiment, one choice problem would be selected at random and implemented, with all equally likely. Any subject who opted to make a purchase in the selected choice problem paid the indicated price out of the participation fee, and was given the item as a snack during the waiting period. Any subject who opted not to make a purchase in the selected choice problem received no snack and retained the entire participation fee.

Group H (2 sessions of 28 subjects each): Subjects considered the same choice problems as in group R, but were aware that all of their decisions were hypothetical, and would not be implemented.

Group M (35 subjects): Subjects considered the same choice problems as in group R, but were told in advance that all but five decisions would be hypothetical. The five real choices were interspersed among the hypothetical choices, but clearly indicated when they were presented. For each subject, the five items were drawn at random from a larger group of fifteen, selected for their representativeness,[73] and each was offered at a price of 75 cents. The purpose of this "mixed" group is to investigate the concern that the low probability with which any given choice problem was implemented in group R influenced purchase frequencies (e.g., if subjects treated the "real" choices as hypothetical).

Group HCT (28 subjects): Subjects performed that same task as in group H, but a "cheap talk" script (as in Cummings and Taylor, 1999) was added to the experimental instructions, with the objective of inducing subjects to take the hypothetical choices more seriously, and thereby minimize hypothetical bias.[74]

Group HL (28 subjects): Subjects performed the same task as in group H, but the questions were modified to elicit the likelihood that the subject would buy the item using a five-point scale (1="very likely," 3="uncertain," 5="very unlikely"), rather than a yes/no decision. The object of this group is to collect information that permits us to distinguish between statements about which subjects are reasonably certain, and those about which they are uncertain, analogously to Champ et al. (1997).

Group HV (28 subjects): Subjects performed the same task as in group HL, except they were asked to indicate how they thought a typical undergraduate of their own gender would answer. The object of these "vicarious" questions is to eliminate image concerns and hence elicit more honest answers, analogously to Rothschild and Wolfers (2011).

Group HWTP (28 subjects): Subjects expressed a hypothetical willingness to pay (WTP) for all of the food items, each of which appeared only once. We employed this protocol because much of the literature explores the accuracy of hypothetical WTPs rather than binary choices. We used the same subjects for groups HWTP and L (below).[75]

Group SWB (28 subjects): For each potential outcome, subjects indicated their anticipated subjective well-being: "How happy would you be if you received this item (and ONLY this item) to eat as a snack during the second part of this experiment, and a price of $X was

[73] Specifically, the distribution of purchase frequencies (among Group R) for the 15 items mirrors the distribution of purchase frequencies for all 189 items.

[74] We would like to 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 developed in Cummings and Taylor (1999).

[75] We combined groups HWTP and L because each required subjects to make fewer responses (i.e., one response for each item, rather than two as in group R and other hypothetical choice groups).

deducted from your show-up payment?" (with 1="very unhappy" and 7="very happy"). Each item appeared twice, once with a price of 25 cents and once with a price of 75 cents.

Group N (28 subjects): Subjects indicated whether each potential outcome would elicit social approval or disapproval: "Imagine that a subject in this experiment paid X cents to eat the item as a snack during the second part of the experiment. Would the typical person approve or disapprove of this purchase?" (with 1="strong disapproval" and 7="strong approval"). These ratings are intended to capture social norms and image concerns.

Group L (28 subjects): Subjects provided liking ratings for each item: "How much would you like to eat this item during the second part of the experiment?" (with 1="not at all" and 7="very much"). Liking ratings are known to be correlated with choices. As noted above, we used the same subjects for groups L and HWTP.

Group S (29-38 subjects):[76] Subjects answered some or all of the following additional questions concerning the food items (answers scaled 1-5): 1) "How much would you later regret eating this snack?" 2) "How tempting is this item?" 3) "If you had no concerns about diet or health, how much would you enjoy eating this item?" 4) "Is this item generally good or generally bad for you?" 5) "Would others form a positive or negative impression of you if they saw you eating this snack?" 6) "Are people likely to understate or overstate their inclination to pick this snack?" The responses to these questions may be useful for predicting choices because each question potentially measures factors related to the degree of hypothetical bias. Questions 1 through 4 address the degree to which immediate gratification conflicts with longer term considerations: we conjectured that hypothetical choices will be more sensitive to long-term costs, and less sensitive to immediate gratification, than real choices. Question 5 addresses concerns for social image: we conjectured that hypothetical choices will be more sensitive to image concerns than real choices. Finally, question 6 may determine whether subjects can provide subjective assessments of hypothetical bias that would be useful for the purpose of predicting choices, even if the sources of the bias remain unclear.

E.2 List of detailed hypothetical evaluations

Detailed hypothetical evaluations include, first, a set of price-specific variables:

- the fraction of respondents choosing purchase in the hypothetical choice question

[76] We collected 29 subject responses to questions 1, 5, and 6, and either 38 or 31 subject responses (depending on the item) to questions 2, 3, and 4.The variation in sample sizes across items for questions 2, 3, and 4, which occurred because of the manner in which the experiment evolved, is not ideal, but we doubt it has a meaningful impact on our results. Initially we collected responses to questions 1, 5, and 6 from a group of 9 subjects, and responses to questions 2, 3, and 4 from a group of 16 subjects, but concerning only 120 of the 189 items. We then collected responses to questions 1, 5, and 6 from a group of 20 subjects, and responses to questions 2, 3, and 4 from a group of 22 subjects, concerning all 189 items. We then collected responses to all six questions from a group of 9 subjects, but only for the 69 items for which we collected no data from the first two groups.

- the fraction of respondents choosing purchase in the hypothetical choice question following the cheap talk script
- the average reported likelihood of purchasing (on a 5 point scale)
- the fraction of respondents stating a likelihood of at least each level (except for "very unlikely" because all respondents choose at least "very unlikely")
- the average vicarious choice likelihood (on a 5 point scale)
- the fraction of respondents stating a vicarious likelihood of at least each level (except for "very unlikely").

Second, variables that are not price-specific; for each of the six questions of Group S (see Appendix E.1; an additional 6×5 variables):

- the average response
- the fraction choosing at least 2, 3, 4, or 5 (ordered such that 5 is most desirable)

Finally, we include the average response for each of the questions asked of Groups SWB, N, and L. For simulations with random treatment assignment, we also include the fraction of respondents whose WTP exceeds the price. In total, this generates 45 or 46 base variables.

E.3 Assessing whether respondents take the "real choice" seriously

We added a "mixed" group, in which subjects were told that five of their choices would be real (that is, one of the five would be chosen at random and implemented), and the rest would be hypothetical. The real choices were clearly identified and interspersed among the hypothetical ones. In that group, the implementation probability for each real choice was 1 in 5 rather than 1 in 378. We elicited 175 real choices through this "mixed" group, pertaining to 15 distinct items (at a price of $0.75). We then pooled that data with 450 choices involving the same 15 items from the "real choice" group, and estimated a logit regression relating the purchase decision to a set of 15 product dummies as well as a "mixed choice group" dummy. If the "real choice group" subjects viewed their choices as real, the coefficient for the "mixed choice group" dummy should be zero; if they viewed those choices as partially hypothetical, then the "mixed choice group" coefficient should be negative given the documented direction of hypothetical bias. In fact, it was positive 0.11, with a standard error of 0.21 (assuming independent observations). The difference is both statistically insignificant and of an economically small magnitude (average marginal effect of less than 2 percentage points). The coefficient indicates that the purchase frequencies were, if anything, slightly higher for real choices in the "mixed choice" group than in the "real choice" group,

which is inconsistent with the hypothesis that participants in the "real choice" group were more inclined to view their choices as hypothetical than were participants in the "mixed choice" group.

E.4 Quantifying "hypothetical noise"

To determine whether hypothetical purchase frequencies, absent sampling uncertainty, are inherently more dispersed across items than real purchase frequencies, we perform the following calculation. For ease of notation, consider all items at a single price.

The observed average hypothetical choice is $H_j = \frac{1}{N}\sum_{i=1}^{N} H_{ij}$ where N is the number of subjects.

The population hypothetical purchase frequency of item j is defined as $\mu_j = \mathbb{E}(H_{ij})$ where the expectation is taken over subjects holding fixed item j, under random sampling of subjects. Denote the average across items of the the population hypothetical purchase frequencies by $\mu = \mathbb{E}(\mu_j)$.

We are interested in $\sigma_H^2 = \mathrm{var}(\mu_j)$ across items j to measure the dispersion of population hypothetical choice frequencies across items.

The sample variance of H_j across items j is $s_H^2 = \frac{1}{J-1}\sum_{j=1}^{J}\left(H_j - \bar{H}\right)^2$ where $\bar{H} = \frac{1}{J}\sum_{j=1}^{J} H_j$ and J denotes the number of items in the sample. Treating both the selection of items and the choice of subjects as random, and allowing for the possibility that the choices of a randomly selected subject may be correlated across items, one can show that

$$\mathbb{E}(s_H^2) = \sigma_H^2 + \sigma_\omega^2(1 - \rho_H)$$

where σ_ω^2 denotes the variance of the sampling error $\omega_j = H_j - \mu_j$ across items j, and ρ_H is the correlation between the sampling errors of two randomly selected items.

Rearranging, we have

$$\sigma_H^2 = \mathbb{E}(s_H^2) - \sigma_\omega^2(1 - \rho_H)$$

To bound σ_ω^2, note that by the law of total variance $\sigma_\omega^2 = \mathrm{var}(\omega_j) = \mathrm{var}(\mathbb{E}(\omega_j|\mu_j)) + \mathbb{E}(\mathrm{var}(\omega_j|\mu_j))$. The conditional expectation in the first term is 0 because $\mathbb{E}(H_j|\mu_j) = \mu_j$. For the second term, note that for any given μ_j, $N \cdot H_j$ is binomial(μ_j,N), such that the sampling error has variance $\mathrm{var}(\omega_j|\mu_j) = \mu_j(1-\mu_j)/N$. Then, $\mathbb{E}(\mu_j(1-\mu_j)/N) < \mu(1-\mu)/N$ by Jensen's inequality because the expression inside the expectation is concave.

Additionally, $\sigma_\omega^2(1-\rho_H) < \sigma_\omega^2$ as long as ρ_H is positive. The correlation between sampling errors across items is likely positive, e.g., because hungry subjects are more inclined to buy all items.

Then

$$\sigma_H^2 = \mathbb{E}(s_H^2) - \sigma_\omega^2(1-\rho_H) > \mathbb{E}(s_H^2) - \sigma_\omega^2 > \mathbb{E}(s_H^2) - \mathbb{E}(\mu(1-\mu)/N)$$

such that $s_H^2 - \bar{H}(1-\bar{H})/N$ is a reasonable estimate of a bound on σ_H^2.

At the high price $s_H^2 = 0.016$ and $\bar{H} = 0.23$, with $N = 28$, such that we bound $\sigma_H^2 > 0.0095$. At the low price $s_H^2 = 0.022$ and $\bar{H} = 0.39$, with $N = 28$, such that we bound $\sigma_H^2 > 0.013$. Those lower bounds exceed, respectively, $s_Y^2 = 0.0083 > \sigma_Y^2$ and $s_Y^2 = 0.0012 > \sigma_Y^2$ calculated analogously using average real choices Y_j in place of hypothetical choices H_j. Because the variances of average real choices across items, σ_Y^2 for high and low prices, are likely considerably smaller than the latter figures (which include sampling error), we conclude that σ_H^2 likely exceeds σ_Y^2 by a wide margin.

F Microfinance Application

F.1 Validation

The design included several checks to ensure that respondents took the survey seriously. First, we asked respondents for the world population and number of people living in poverty (with free text answers); except for a handful of responses, all answers are reasonable. Second, after reading the instructions, participants responded to two simple questions to validate understanding of the study. In order to complete the study, participants had to respond correctly. Third, after illustrating different features of loan postings, respondents had to answer three further understanding questions about these features (multiple choice with 3 options); 70% answered all questions correctly, and a majority of those answering incorrectly had only one incorrect answer. After answering the understanding questions, respondents were shown one additional screen for each incorrect answer, explaining the correct answer and asking them to answer the remaining questions in the survey more carefully. Fourth, responses to one question were incentivized. Fifth, in the final demographic survey, respondents were asked to rate the following three statements along the same Likert scale ranging from 'Strongly Disagree' to 'Strongly Agree': 'I made each decision in this study carefully', 'I made decisions in this study randomly', and 'I understood what my decisions meant.' A careful respondent should agree with the first and last statement but disagree with the middle; agreement or disagreement with all statements reveals that a respondent made careless decisions. 75% of respondents agreed with the first and last statement, and disagreed with the middle; 56% did so strongly.

F.2 Mean squared error with measurement error in covariates

In Section 4.4, we propose estimating our method using subsamples of hypothetical evaluations only of respondents passing certain thresholds in their predictive quality for other settings. Suppose that, in an infinite sample, we can estimate $\boldsymbol{\beta} = \boldsymbol{\beta}_{0,0} = \boldsymbol{\beta}_{1,1}$ from Assumption 3 by finding the threshold r^* that minimizes mean squared error:

$$r^* = \arg\max_r \mathbb{E}\Big((Y_j - \boldsymbol{H}_j^r \boldsymbol{\beta}^r)^2\Big)$$
$$\implies \boldsymbol{\beta} = \boldsymbol{\beta}^{r^*}$$

where $\boldsymbol{H}_j^r$ are the average evaluations for setting j based on an infinite number of respondents passing threshold r, as well as the intercept.[77]

We estimate the squared error of using threshold r in finite samples as follows.

We use an instrumental variables estimator for $\boldsymbol{\beta}^r$. In finite samples, there may be relatively few responses $\boldsymbol{H}_{kj}$ to aggregate when using a strict correlation threshold r. That would confound the comparison of OLS estimates for different thresholds with differential attenuation bias due to classical measurement error.[78] To avoid such differential biases, we split the respondents into halves, to form aggregates $\boldsymbol{H}_j^{r,A}$ and $\boldsymbol{H}_j^{r,B}$ with independent measurement errors. We then estimate $\boldsymbol{\beta}^r$ by regressing outcomes Y_j on $\boldsymbol{H}_j^{r,A}$, using $\boldsymbol{H}_j^{r,B}$ as instruments (Fuller, 1987). We reverse the use of $\boldsymbol{H}_j^A$ and $\boldsymbol{H}_j^B$ and average the resulting coefficient estimates. We use leave-one-out estimates for $\boldsymbol{\beta}^r$: for setting j, we use all settings except j to compute these instrumental variables estimates, say $\hat{\boldsymbol{\beta}}_{-j}^r$.

To correct the estimate of the mean squared error criterion for the measurement error due to small samples of evaluations for strict thresholds, we compute it as

$$\frac{1}{J}\sum_{j=1}^{J}(Y_j - \boldsymbol{H}_j^{r,A}\hat{\boldsymbol{\beta}}_{-j}^r)(Y_j - \boldsymbol{H}_j^{r,B}\hat{\boldsymbol{\beta}}_{-j}^r) - \frac{1}{J}\sum_{j=1}^{J}(Y_j - \boldsymbol{H}_j^{r,A}\hat{\boldsymbol{\beta}}_{-j}^r)\frac{1}{J}\sum_{j=1}^{J}(Y_j - \boldsymbol{H}_j^{r,B}\hat{\boldsymbol{\beta}}_{-j}^r).$$

The first term computes the squared prediction error for setting j as a product of the errors of the predictions made using $\boldsymbol{H}_j^{r,A}$ and $\boldsymbol{H}_j^{r,B}$. In expectation, $\mathbb{E}(\boldsymbol{H}_j^{r,A}) = \mathbb{E}(\boldsymbol{H}_j^{r,B}) = \mathbb{E}(\boldsymbol{H}_j^r)$ and $\mathbb{E}(\boldsymbol{H}_j^{r,A}\boldsymbol{H}_j^{r,B}) = \mathbb{E}((\boldsymbol{H}_j^r)^2)$ because the measurements are unbiased and independent. Hence, the first term estimates mean squared error. The second term is a small-sample correction that vanishes in large samples. In finite samples, $\frac{1}{J}\sum_{j=1}^{J} Y_j \neq \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{H}_j^{r,A}\hat{\boldsymbol{\beta}}_{-j}^r$ and $\frac{1}{J}\sum_{j=1}^{J} Y_j \neq \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{H}_j^{r,B}\hat{\boldsymbol{\beta}}_{-j}^r$ in part due to the measurement error in $\boldsymbol{H}_j^{r,A}$ and $\boldsymbol{H}_j^{r,B}$.

[77] In principle, one could microfound a different criterion for selecting r^* than mean squared error. In our applications, we find that the particular criterion used does not have substantial effects on the estimate if we use approximate residual balancing; intuitively, that method guards against selecting incorrect thresholds in finite samples.

[78] With multiple regressors, the bias in their coefficient estimates due to measurement error could go in any direction.

The second term removes the effect of this error on the estimated mean squared error.