CAUSAL INFERENCE FROM HYPOTHETICAL EVALUATIONS

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

December 9, 2021

Abstract

This paper explores methods for inferring the causal effects of treatments on choices by combining data on real choices with hypothetical evaluations. We propose a class of estimators, identify conditions under which they yield consistent estimates, and derive their asymptotic distributions. The approach is applicable in settings where standard methods cannot be used (e.g., due to the absence of helpful instruments, or because the treatment has not been implemented). It can recover heterogeneous treatment effects more comprehensively, and can improve precision. We provide proof of concept using data generated in a laboratory experiment and through a field application.

KEYWORDS: causal inference, hypotheticals, counterfactuals, machine learning

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

The problem of causal inference is central in empirical social science. This paper considers the canonical task of inferring the effects of a treatment, such as a price or a policy intervention, on choices.[1] Two classes of challenges arise in these settings. First, because the treatment of interest often results from human decision making, it is potentially endogenous, and may therefore bear a spurious relation to choice. Second, the treatment may be rare. For example, it may be an innovative policy adopted by a single jurisdiction. In such circumstances, there is little or no opportunity to observe its effects and to distinguish them statistically from random variation. Even more challenging, the treatment may simply be a proposal that remains untested.

With respect to the first set of challenges, a common approach is to infer causal relationships by focusing on variation in the treatment arising from arguably exogenous factors (instruments), or from discontinuities. Unfortunately, in some applications, suitable instruments are difficult to find. Even when they are available, estimates of the causal relationship may be imprecise, particularly if the connection between the treatment and the instruments is relatively weak. Moreover, when responses to the treatment are heterogeneous, these methods may not identify the particular causal effects that are of interest to the analyst. Nor do they offer solutions to the second set of challenges.

An alternative is to ask people, hypothetically, what they would choose under various conditions, as in the literature on 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 would provide a simple solution to both challenges, because it does not rely on observed treatments.[2] In principle, it would even allow the analyst to recover treatment effects for arbitrary subsets of the population. Unfortunately, there is strong evidence that 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.

Our objective in this paper is to explore ways to address the challenges of causal inference by exploiting the information contained in answers to hypothetical questions. Our strategy

[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 6.2. 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."

is to combine hypothetical responses (in which treatment is unconfounded but measures of outcomes may be biased) with observational data on choices (in which treatment may be confounded but real outcomes are measured without bias). We consider not only subjective "aggregators" such as stated preference and hypothetical choices, but also a variety of hypothetical responses that capture underlying motivations (such as temptation or social image), which may relate to the magnitude and direction of hypothetical bias.

In essence, we propose estimating a predictive relationship between hypothetical responses and real choices in observational data, and then using that relationship to infer the effects of counterfactuals. To be more specific, consider an environment in which the treatment of interest, $w \in \{0, 1\}$, varies across settings, indexed by j. Examples include prices varying over 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 effect of treatment, $\tau = \mathbb{E}(Y_j(1) - Y_j(0))$. 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 data pertaining to setting j, $\boldsymbol{H}_j(w)$, for both treatment states. First. we train a model to predict outcomes in the realized treatment states, $Y_j(W_j)$, based on the corresponding hypothetical responses, $\boldsymbol{H}_j(W_j)$.[4] For the linear case, we estimate the model $Y_j(W_j) = \boldsymbol{H}_j(W_j)\boldsymbol{\beta}$ and recover the coefficient vector $\hat{\boldsymbol{\beta}}$. Second, we use that relationship to predict the outcome for each treatment state. The difference yields an estimate of the treatment effect, $\hat{\tau}_{\text{p}} = \big(\overline{\boldsymbol{H}(1)} - \overline{\boldsymbol{H}(0)}\big)\hat{\boldsymbol{\beta}}$. We develop a simple linear estimator which is 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.[5]

As long as the predictive relationship is stable, this method should yield unbiased estimates of treatment effects. In effect, we rely on the estimated prediction equation to unwind the systematic biases embedded in the hypothetical responses. We do not, however, claim that stability of the predictive relationship is generally guaranteed. On the contrary, our contribution is to (i) articulate conditions that would yield stability, thereby clarifying the contexts to which 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 estimates of treatment effects that are close to ground-truth experimental

[4] There is an analogy between this approach and methods used in the literature on demand estimation, which has made progress by decomposing demand for products (product space), onto their physical characteristics Lancaster (characteristic space, 1966). In essence, we further treat underlying motivations as characteristics and elicit them through survey responses.

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

estimates, even under conditions that render standard methods inapplicable.[6]

Our laboratory setting estimates the demand for various snacks as a function of prices. We ask some participants to make real purchase decisions for each snack at two prices, $0.25 and $0.75, which allows us to determine ground truth. We ask other participants to evaluate each snack-price pair hypothetically along several dimensions. We simulate a data set with endogenous variation by assuming that each snack is offered only at a single price, correlated with demand, as well as data sets with no price variation.

Our field setting assesses the effects of matching provisions on lending through a microfinance crowdfunding platform. The observational data include the presence of a match on each borrower profile and the speed at which it attracted funding. We gathered hypothetical data by asking Amazon Mechanical Turk workers to assess these profiles in both the unmatched and matched states. We determined ground truth by working with the platform to implement a field experiment that randomly varied the match status of profiles.

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

First, our method can recover average treatment effects even in settings where standard methods are inapplicable. In both applications, the difference between treated and untreated units yields severely biased estimates of the treatment effect due to endogenous assignment of the treatment. 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 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 our lab setting, these alternative protocols instead simply introduce additional biases that in most cases neither reduce nor eliminate 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 where there is no variation in the treatment. While standard observational methods are inapplicable in these cases,

[6]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. 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.

it is entirely possible that two disjoint collections of choice problems induce overlapping distributions of motivational responses. In that case, the relationship between choices and hypothetical reactions in one treatment state potentially applies in the alternative state. For our laboratory data, we find that the distribution of subjective responses for the high price largely spans the distribution of subjective responses for the low price. Consequently, if all snacks are observed at the high price, our method recovers estimates very close to the ground truth. However, the distribution of subjective responses for the low price does not span the distribution of subjective responses for the high price nearly as well. If all snacks are observed at the low price, our method still recovers a reasonable estimate, but it is further from the truth, as one would expect.

Third, our method yields more comprehensive measures of heterogeneous treatment effects than standard approaches, and can do so even without exogenous (randomized) treatment variation. While standard methods measure only the local average treatment effects (LATEs) among compliers, our approach allows the analyst to recover treatment effects for arbitrary subgroups. In our laboratory application, we show that measures of heterogeneous treatment effects that condition on observable characteristics capture only a small fraction of the underlying heterogeneity. We exhibit the value of an improved ability to measure response heterogeneity through a price-setting exercise, wherein our method enables the price setter to dramatically increase simulated profits. 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 estimates. 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 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 variation is available, particularly when treatment groups are unbalanced in their sample sizes. Assuming the predictive relationship is stable over treatment states, imbalance does not fundamentally impact the precision with which one estimates it (our first step). Furthermore, because we use hypothetical data in both treatment states to predict outcomes for every setting, imbalance has no impact on the precision of our second step. For these reasons, we obtain precise measures of treatment effects even when the treatment is rare (or not observed) in practice. More generally, the high correlations between real choices and hypothetical reactions tend to deliver a high level of precision.

To be clear, we do not offer this method as a panacea, nor do we recommend its indiscrim-

inate application. The method is more likely to produce accurate results when settings can be described or depicted comprehensively to survey respondents, when respondents are better equipped to visualize actual decisions, and when it is possible to select survey respondents who resemble, but are distinct from, the impacted decision makers. Nonetheless, in some settings the approach may provide a reasonably 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 introduces a conceptual framework to outline the conditions necessary for the method. Section 3 describes the estimator and its properties. Section 4 describes the lab setting and its results, and Section 5 the field setting. Section 6 outlines extensions, and Section 7 concludes.

2 Conceptual Framework

2.1 Characteristics of intended applications

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.

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 $\boldsymbol{\xi}_{ij} \sim F_j^{typ}$ that may vary across individuals. Thus, $W_j = W_j(\boldsymbol{X}_j, \boldsymbol{\eta}_j, F_j^{typ})$.

Individual i's choice in setting j depends on the treatment, stable characteristics of the setting, $\boldsymbol{X}_j$ and $\boldsymbol{\eta}_j$, and *realized* conditions $\boldsymbol{\xi}_{ij} \sim F_j$ that i experiences in setting j. Thus, $Y_{ij} = Y(W_j, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$.[7] 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 $\boldsymbol{\xi}_{ij}$ 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.

[7] 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.

The average outcome in setting j with treatment state w is

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

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

$$W_j \perp\!\!\!\perp \{\epsilon_j(w)\}_{w \in \mathbb{W}} \mid \boldsymbol{X}_j, \boldsymbol{\eta}_j, F_j^{typ}.$$

We offer four concrete example applications to fix ideas:

Product demand. The analyst seeks to estimate price elasticities for products falling within some category (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. Our laboratory experiment simulates this application.[8]

Matching of charitable contributions. The analyst wishes to determine the effects of offering matching contributions for charitable donations to appeals posted on an online platform, accounting for the fact that sponsors choose the appeals for which matches are available (Karlan and List, 2007; Huck and Rasul, 2011). Here, settings correspond to appeals, the treatment is the existence of a match, and the outcomes are donation decisions by the platform's users. Our field experiment resembles this application.

401(k) plans and saving. The analyst intends to estimate the effects of pension plans on retirement saving, accounting for the fact that employers take workers' preferences into consideration when deciding whether to offer such plans (Engen et al., 1996; Poterba et al., 1996). Here, settings correspond to employers, the treatment is the existence of a 401(k) plan, and the outcomes are workers' saving decisions.

Default options for organ donations. Driver's license application forms commonly include an option to register as an organ donor. The analyst hopes to evaluate the effect of the

[8] 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.)

default option (opt-in versus opt-out) on organ donation elections, accounting for the fact that lawmakers may set the default based in part on the general inclinations of the electorate (Kessler and Roth, 2012, 2014). Here, settings correspond to states, the treatment is the default option, and the outcomes are applicants' organ donation elections.

Note that, in each of these examples, the actual chosen outcomes depend in part on conditions that materialize after treatment is determined.

2.2 Method of causal inference

2.2.1 The basic idea

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

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

We assume that these motivations reflect the treatment as well as the observed and unobserved characteristics of the individual and the setting: $\boldsymbol{Q}_{ij}(w) = \boldsymbol{Q}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$. At this level of generality, external conditions, including the 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

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

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

In practice, the credibility of the exclusion restriction depends on the measurement of $\boldsymbol{Q}_{ij}(w)$. We propose collecting variables , $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(w)$, describing the marginal distributions $F_j^{typ,\boldsymbol{Q}(w)}$ (such as moments and percentiles) for each observed setting j, under typical conditions in both treatment states. For simplicity, we focus here on the case of binary treatments, $W_j \in \{0, 1\}$.

Suppose we also observed the potential outcomes $Y_j^{typ}(w)$ under typical conditions in both treatment states. We could then regress $Y_j^{typ}(w)$ on these distributional characteristics $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(w)$, pooling observations from all settings and treatment conditions, and then use the estimated equations to compute predicted choices, $\hat{Y}_j(0)$ and $\hat{Y}_j(1)$. As long as one selects 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 be equal to the predicted treatment effect, $\hat{Y}_j(1) - \hat{Y}_j(0)$. With multi-valued treatments, one could

similarly predict 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.

Now imagine that we instead observe only $Y_j(W_j)$, the choices in setting j under realized rather than typical conditions, and only for the treatment condition that actually prevails. Then we could run the same regression using the available data (i.e., regress $Y_j(W_j)$ on $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(W_j)$), and use it to construct $\hat{Y}_j(1) - \hat{Y}_j(0)$ exactly as before. If the distributions of the covariates $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(0)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(1)$ 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 $\boldsymbol{Q}_{ij}(W_j)$, which 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 $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(0)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(1)$ are *balancing scores* in the sense of Rosenbaum and Rubin (1983).[9]

The other difference between this procedure and the original is that it employs 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 $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(W_j)$ 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 $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(W_j)$. 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, which we motivate below: if $\epsilon_j(W_j)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{Q}}(W_j)$ were correlated, then presumably F_j^{typ} does not represent the most typical condition.

It follows that the differences between the original and alternative procedures are innocuous under reasonable assumptions. The requirements of the method therefore largely boil down to whether it is possible to measure motivations sufficiently well.

2.2.2 The measurement of motivations

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). We use a distinct sample to avoid real choices contaminating hypothetical evaluations, or vice versa.[10] 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

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

[10] If hypothetical evaluator k has already experienced setting j in realized treatment state W_j, their hypothetical responses for that state, $\boldsymbol{H}_{kj}(W_j)$, may be close to the truth. However, their hypothetical responses for the

otherwise the same. The validity of this approach depends on how hypothetical motivations for survey respondents relate to typical motivations for decision makers. Here we identify six potential concerns and explain how we address them. Our applications provide proof of concept that our solutions can suffice with real data.

Hypothetical biases. Although hypothetical responses have been found to be systematically biased (List and Gallet, 2001; Little and Berrens, 2004; Murphy et al., 2005), our method is not vulnerable to these biases because conditioning on a variable is equivalent to conditioning on a monotonic transformation (biased measure) of that variable. That said, including measures of motivations that relate to the directions and magnitudes of such biases may increase precision.

Comprehensiveness of elicited motivations. The omission of important motivations is potentially problematic. However, it is not necessary to include separate measures of every pertinent motivation. It suffices that the elements of $\boldsymbol{H}_{kj}(w)$ collectively span the empirically significant motivations. Eliciting motivational aggregates (e.g., hypothetical or vicarious choice, anticipated satisfaction) helps us achieve this objective without enumerating all the components. Using multiple aggregates along with measures of the most relevant motivational factors allows the regression, in effect, to re-weight the underlying components. To the extent the set of basic human motivations is limited (as suggested, e.g., by Maslow, 1943), spanning may be relatively easy to achieve. As a diagnostic, one can add measures of component motivations sequentially starting with those presumed most pertinent to the application, and check the stability of the estimated treatment effect.[11]

Overlap of elicited motivations. More generally, for some decision problems and treatment states, basic motivators $\boldsymbol{Q}_j(w)$ may be "outside the convex hull" of those observed in the real data, $(Y_j, \boldsymbol{Q}_j(W_j))$. In these cases, we would be concerned about the extrapolation required of our method. Overlap is amenable to empirical investigation, as we demonstrate in our applications.

Comprehensiveness of descriptions. The method assumes that a survey can convey $\boldsymbol{\eta}_j$, even though the econometrician cannot distill or observe it. This assumption is plausible when the econometrician's challenge is coding pertinent aspects of the choice setting, which can be presented naturally to respondents. For example, when survey respondents view an image of a Snickers bar, they presumably have the same nuanced understanding of its features as when

unrealized treatment state, $\boldsymbol{H}_{kj}(1 - W_j)$ may be subject to different biases. Those responses could be anchored by the real state, and thus fail to account for how treatment affects the outcome. Or they could overstate the effect of treatment if the hypothetical makes treatment more salient than it would be in real decisionmaking (e.g. "how much would you buy if sales tax was $x\%$"?). Although our method can correct biases that affect real and counterfactual treatment states symmetrically, these biases asymmetrically affect the states and thus would not be corrected.

[11] This procedure is similar to demonstrating stability across increasingly complex specifications in observational analyses.

making actual purchase decisions. Similarly, respondents in our microlending application who view actual loan solicitation postings can process the complex context of the treatment. The same is true for a respondent who views the entire driver's license application in the organ donation application. However, our method is less applicable to complex decision problems that are difficult to depict or describe comprehensively.[12] Similarly, hypotheticals are most useful when counterfactuals fall within the respondents' normal experience. They may be of more limited usefulness if an unobserved treatment state is hard for individuals to imagine, or the corresponding real choice would induce significantly more reoptimization (for example, if I decided to take public transit instead of driving, that may induce changes to other choices that I may not foresee when asked hypothetically). For applications of the estimation of demand, prices may be easier for respondents to evaluate than novel combinations of characteristics.[13]

Noisy measurement of motivations. Because we calculate the summary statistics $\boldsymbol{D}_j^{typ,\boldsymbol{H}}$ based on a sample of survey respondents, they are subject to sampling error. Any tendency to answer hypothetical questions based on idiosyncratic rather than typical conditions contributes to this noise.[14] The analyst can reduce noise by increasing the size of the sample, or correct for it using various methods for addressing classical measurement error in regressors (see Section 6.1).

Similarity of populations. The method assumes that survey respondents are representative of the actual decision makers in terms of their responses to hypotheticals. When it is not easy to draw respondents from the same population as decision makers, several strategies are available. First, one can select respondents with representative characteristics, such as demographics, interests, and familiarity with the choice setting. Second, one can ask respondents to envision the choice of a typical decision maker vicariously. Third, one can include only the subset of respondents whose hypothetical responses are sufficiently close to actual choices for observed settings (we demonstrate this approach in Section 6.1). The method also assumes that the decision-making populations for different settings exhibit a similar action-motivation relationship, even if tastes (which determine motivations) differ. This assumption is automatically satisfied in applications for which all settings pertain to the same population, such as the examples involving product demand and charitable contributions mentioned above. It may also be reasonable for applications in which settings pertain to broad geographic areas, such as U.S. states. Similarity of populations is difficult to

[12] For example, a consumer selecting an automobile to purchase may gather ratings, take test drives, and have conversations with friends; it would be difficult to present hypothetical respondents with the same factors.

[13] However, this may break down at extremes: hypothetical evaluations may not anticipate how I may use a dramatically cheaper good in novel ways, or seek novel substitutes for one that is dramatically more expensive.

[14] The most problematic case arises when the conditions respondents visualize are both non-typical and correlated, e.g., because they project conditions at the time of the survey.

justify when each setting corresponds to a distinct individual, and we do not recommend using our method to study those applications.

2.2.3 Potential advantages

In suitable applications, our approach can in principle accurately estimate treatment effects, even when there is no exogenous variation in treatment. Our method can potentially address other challenges in causal inference:

Unobserved treatments. When a treatment is merely a proposal, so that there is no observed variation in the treatment state W_j, conventional methods for measuring treatment effects are inapplicable. In contrast, our method is easily applied. The central insight is that new treatments do not ordinarily create new motivations. As long as the treatment does not alter the relationship between choices and motivations, one can use that relationship to project choices in the unobserved treatment regime. That extrapolation will be particularly reliable when the motivational responses for the two treatment regimes have overlapping distributions.

Precision. Conventional methods of causal inference may produce imprecise estimates of treatment effects for at least two reasons. First, either the treatment or its absence may be relatively uncommon. In such cases, even when the treatment states are random, comparisons of average outcomes tend to have large standard errors. Second, observational methods isolate consideration to the component of variation in treatment that is plausibly exogenous; when this component is small, estimates of treatment effects will tend to be imprecise.

Our approach potentially avoids these difficulties because we elicit both treated and untreated hypothetical intent (along with other hypothetical responses) for every setting. As a result, our estimate of the ATE, $\frac{1}{J}\sum_{j=1}^{J}\left[\hat{Y}_j(1) - \hat{Y}_j(0)\right]$, includes all settings and is always based on the same number of observations for both treatment states. Furthermore, as long as the variation in $\boldsymbol{Q}$ remains well within-sample for the scarce treatment, the precision of predictions will be comparable for the scarce and abundant treatment states.

Measurement of heterogeneous treatment effects. Treatment effects commonly vary across units (here, across settings). Standard observational methods identify a Local Average Treatment Effect (LATE) among the subset of units that change their choices in response to the instrument or discontinuity (the compliers; Imbens and Angrist, 1994). They do not permit the analyst to recover the overall Average Treatment Effect (ATE), or ATEs for arbitrary subgroups. In contrast, our approach permits the analyst to predict the ATE overall or the ATE for any subgroup with specified characteristics. Naturally, the estimate will tend to be more precise for groups containing more settings.

Complexly heterogeneous treatments. Treatments may vary in complex ways that are not easily described by a small collection of measured characteristics. As an example, consider the case of organ donation elections on drivers' license applications. The literature focuses on a single dimension of the treatment: whether the nature of the election is opt-in or opt-out (Kessler and Roth, 2012, 2014). Yet the actual treatments differ in many other ways, including the wording of the organ donation question (which in some cases includes explanatory text), the placement of the question on the form, the color and size of the font, and other aspects of the form. Organ donation elections may respond in complex ways to all of these characteristics and their interactions. The conventional approach requires the analyst to either assume these effects away, treat them as orthogonal to the effect of interest, or model the heterogeneity of the treatment. The last option can be challenging. For example, dummy variables can capture only the coarsest features of wording.

In such settings, our approach permits one to distill the effect of interest, for example by eliciting hypothetical responses to two versions of each drivers' license application form: the one that is actually used, and one that alters the pertinent question from opt-in to opt-out (or vice versa) while preserving all other features. The difference in predicted responses then measures the impact of this limited change conditional on the other features, whatever they are. In principle, one can also use our method to explore the effects of more elaborate and complex alterations, such as changes in wording: simply assess hypothetical responses to the design of interest, and then use the same predictive relationship to infer its effects.

3 Estimation

In this section, we identify statistical assumptions that justify the proposed estimators, and describe their asymptotic distributions. We focus on a set of estimators that employ data on actual outcomes, the actual assignments of a binary treatment, and predictors for each treatment state, $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1))_{j=1}^{J}$. The vector of predictors, $\boldsymbol{H}$, includes $\boldsymbol{D}^{typ,\boldsymbol{Q}}$, the relevant features of the distribution of hypothetical evaluations; they may also include measures of the setting's fixed characteristics $\boldsymbol{X}$.

The basic estimation strategy employs the following two-step procedure:

Step 1. Using data pertaining to the realized treatment states, estimate the parameters $\boldsymbol{\beta}$ determining the relationship between the realized outcome and the predictors:

$$Y_j = \mu(\boldsymbol{H}_j(W_j), \boldsymbol{\beta}) + \epsilon_j$$

Step 2. Using the estimated relationship from Step 1, predict outcomes for both states, and

take the difference:

$$\hat{\tau}_{\text{p}} = \frac{1}{J}\sum_{j=1}^{J}\Big(\hat{Y}_j(1) - \hat{Y}_j(0)\Big) = \frac{1}{J}\sum_{j=1}^{J}\Big(\mu(\boldsymbol{H}_j(1), \hat{\boldsymbol{\beta}}) - \mu(\boldsymbol{H}_j(0), \hat{\boldsymbol{\beta}})\Big)$$

The function μ encompasses the relationship $Y^*(\boldsymbol{Q})$ from Section 2, but in the interests of generality it also accommodates the possibility that this relationship depends on each setting's fixed characteristics. The estimation error ϵ reflects the difference between typical conditions and realized conditions. For the time being, we assume that our measures of hypothetical evaluations are based on arbitrarily large populations of survey respondents; we address issues arising from the use of finite survey samples at the end of the section. We begin by studying an estimator for the case in which μ is linear, and then develop a machine learning estimator for more general specifications. An accompanying R package is available on Github: `https://github.com/michaelpollmann/hypeRest`.

3.1 A simple linear estimator

The simplest strategy is to approximate μ with a linear specification, potentially including interactions. Approximate linearity is especially plausible when $\boldsymbol{H}_j(w)$ includes hypothetical choices, as long as systematic hypothetical choice biases are reasonably simple. For this case, we proceed as follows:

Step 1. Use OLS to regress the realized outcome on predictors for the realized treatment state $\boldsymbol{H}_j = \boldsymbol{H}_j(W_j)$:

$$Y_j = \boldsymbol{H}_j\boldsymbol{\beta} + \epsilon_j.$$

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

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

where $\overline{\boldsymbol{H}(w)} = \frac{1}{J}\sum_{j=1}^{J} H_j(w)$ is the sample average of the predictors under treatment state $w \in \{0, 1\}$ for all decision problems.

3.1.1 Assumptions

Our first assumption, unconfoundedness, holds that variation in the treatment and the outcome are independent, conditional on the predictors. As we discussed in Section 2, this assumption is reasonable if the treatment assignment is based on typical conditions as opposed to the specific conditions that actually materialize, and if the predictors include hypothetical evaluations that span the set of motivations.

Assumption 1. Unconfoundedness. *Treatment assignment is unconfounded conditional on hypothetical evaluations*:

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

As discussed in the previous section, we hypothesize that basic motivations capture all information relevant for choice, including the treatment state. Accordingly, if two settings or treatment states induce the same motivations under typical conditions, they should yield the same expected outcomes. Assuming the predictors span the basic motivations, this *exclusion restriction* justifies the following two assumptions:

Assumption 2. State specific hypothetical evaluations. *For the potential outcome $Y_j(w)$ in treatment state $w \in \{0, 1\}$, only the predictors for that treatment state, $\boldsymbol{H}_j(w)$, matter:*

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

Assumption 3. Invariant mapping. *The mapping between potential outcomes and predictors is the same irrespective of the treatment state:*

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

Our framework attributes deviations between observed outcomes and these conditional expectations to the fact that hypothetical responses envision typical conditions, while actual choices reflect realized conditions.

Our final assumption (which we state without imposing the last two) introduces linearity.

Assumption 4. Linearity. *The conditional expectations of potential outcomes are linear in the predictors:*

$$\mathbb{E}\Big(Y_j(0) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}_1, \boldsymbol{H}_j(0) = \boldsymbol{h}_0\Big) = \boldsymbol{h}_1\boldsymbol{\beta}_{0,1} + \boldsymbol{h}_0\boldsymbol{\beta}_{0,0}$$
$$\mathbb{E}\Big(Y_j(1) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}_1, \boldsymbol{H}_j(0) = \boldsymbol{h}_0\Big) = \boldsymbol{h}_1\boldsymbol{\beta}_{1,1} + \boldsymbol{h}_0\boldsymbol{\beta}_{1,0}$$

In practice, to make linearity more plausible, one may wish to include second order terms and interactions. When this strategy generates a large number of covariates, a high-dimensional estimator may be more appropriate, which we introduce later.

3.1.2 Asymptotic distribution

The following theorem characterizes the asymptotic distribution for our linear estimator.

Theorem 1. *Suppose the data* $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1))_{j=1}^J$ *are a random sample of independent observations. Under Assumptions 1, 2, 3, and 4, as well as the standard regularity conditions, the parametric estimator* $\hat{\tau}_{\text{p}}$ *is consistent for the average treatment effect and asymptotically normal with*

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

where

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

$\boldsymbol{V}^{\text{ols}} = \mathbb{E}\Big(\boldsymbol{H}_j'\boldsymbol{H}_j\Big)^{-1}\mathbb{E}\Big(\boldsymbol{H}_j'\boldsymbol{H}_j(y - \boldsymbol{H}_j\boldsymbol{\beta}^*)^2\Big)\mathbb{E}\Big(\boldsymbol{H}_j'\boldsymbol{H}_j\Big)^{-1}$ *is the asymptotic variance matrix of the OLS estimator from step 1,* $\boldsymbol{H}_j = \boldsymbol{H}_j(W_j)$*, and* $\boldsymbol{\beta}^* = \boldsymbol{\beta}_{0,0} = \boldsymbol{\beta}_{1,1}$ *from Assumption 4.*

Proof: The result follows by writing the two-step estimator in the GMM framework (cf. Newey and McFadden, 1994); see Appendix C.1 for details.

Estimating the asymptotic variance matrices, and obtaining standard errors, is straightforward: one replaces expectations with sample moments and substitutes the step 1 and step 2 estimates for the unknown parameters. The variance in Theorem 1 accounts for sampling of settings and their characteristics from a super-population.[15]

3.1.3 Discussion

Our estimators and results differ from those associated with standard methods that exploit unconfoundedness: we do not use the treatment indicator directly, but instead use hypothetical covariates describing the outcome under *both* treatment states, and in step 1 we pool all observations irrespective of treatment status. This procedure allows us to estimate treatment effects even when there is no variation in treatment assignment, and can drastically reduce variance.

[15]The first term in V_τ is the naive variance of $\hat{\tau}_{\text{p}}$ if one takes the step 1 coefficient estimates as given, considering only variance in the second-step arising through sampling of the difference in hypothetical evaluations $(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))$. The second term in V_τ is the variance of $\hat{\tau}_{\text{p}}$ if one instead considers hypothetical evaluations of the settings as given, allowing application of the Delta method to the step 1 coefficient estimates. The third term adjusts for the (properly scaled) covariance between the OLS moments of step 1, $\boldsymbol{H}_j'(Y_j - \boldsymbol{H}_j\boldsymbol{\beta}^*)$, and treatment effect prediction moment of step 2, $\tau^* - (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\boldsymbol{\beta}^*$. One should use an estimator of the variance of the OLS coefficients, $\boldsymbol{V}^{\text{ols}}$, that is robust to heteroskedasticity.

We assume that motivations capture all the effects of the treatment on the outcome of interest, in a manner similar to statistical surrogates (for instance, Prentice, 1989; Begg and Leung, 2000; Frangakis and Rubin, 2002; Athey et al., 2020). However, statistical surrogates are observed only under the realized treatment state, whereas we observe hypothetical evaluations under both (all) treatment states; this consideration leads to different estimators and properties.

The theoretical results treat the hypothetical evaluations contained in $\boldsymbol{H}_j(0)$ and $\boldsymbol{H}_j(1)$ as given (population statistics), rather than as aggregations of finite samples. This is appropriate when there are many respondents relative to the number of settings, so that the variation from sampling respondents is not of first-order importance. In practice, it is often easier to increase the number of individuals surveyed than to sample additional settings. As an alternative, one can think of our analysis as conditional on the finite sample of individuals from whom we elicit hypothetical evaluations.[16]

3.2 Estimators for high-dimensional evaluations and non-linear relationships

Machine learning estimators that perform selection and shrinkage may outperform the linear estimator if the sample is large, hypothetical biases are complicated, or there are many types of hypothetical evaluations, most of which may have only limited predictive power. We develop such an estimator for cases involving linearity in high-dimensional hypothetical evaluations. In Appendix C.2, we show that our approach is not generally tied to any parametric assumptions, because treatment effects are identified non-parametrically. In Appendix C.3, we provide a doubly robust moment condition for estimation using arbitrary machine learning methods.[17]

Let $\boldsymbol{Z}_j(w) = g(\boldsymbol{H}_j(w))$ be the covariate vector for setting j, including predictors $\boldsymbol{H}_j(w)$ for treatment state $w \in \{0, 1\}$, 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:

[16] A potential concern then arises because the unconfoundedness assumption is based on motivational responses among a much larger population. In that case, we can treat the discrepancy as a source of measurement error, and apply standard corrections. We describe one such correction in Section 6.1.

[17] Interestingly, estimators based on the doubly robust moment condition do not share all the advantages of those discussed in the main text (see Section 6.2), so we do not study them in detail here.

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:

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

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

Step 1b. Compute approximate balancing weights

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

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

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

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 $\boldsymbol{\gamma}^t$ and $\boldsymbol{\gamma}^c$, computed in step 1b, are meant to reduce estimation errors for $\mathbb{E}\big(\mathbb{E}(Y_j(1)|\boldsymbol{H}_j(1))\big)$ and $\mathbb{E}\big(\mathbb{E}(Y_j(0)|\boldsymbol{H}_j(0))\big)$ in the first term of step 2, under the assumption of linearity.[18]

[18] Specifically, the objective functions in step 1b have two parts. Introducing $\|\tilde{\gamma}\|_2^2$ reduces the variance of the estimator by penalizing deviations from equal weights. Introducing $\|\overline{\boldsymbol{Z}(w)} - \boldsymbol{Z}^T\tilde{\boldsymbol{\gamma}}\|_\infty^2$ 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.

3.2.1 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 $\mathbb{H}_0$ and $\mathbb{H}_1$ are the supports of the distributions of predictors in the control and treatment states, respectively, then for all $\boldsymbol{h} \in (\mathbb{H}_0 \cup \mathbb{H}_1)$, we have for some $\eta > 0$ at least one of*

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

or

$$\eta < \Pr(W_j = 1 \mid \boldsymbol{H}_j(1) = \boldsymbol{h})$$

A sufficient condition for this assumption is that, for any value of the predictors $\boldsymbol{h} \in (\mathbb{H}_0 \cup \mathbb{H}_1)$, we observe (a growing number of) settings j for which the hypothetical evaluations corresponding to the realized treatment state coincide with h, i.e., $\boldsymbol{H}_j(W_j) = \boldsymbol{h}$. 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 1) 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, 4 (here linearity in high-dimensional covariates $\boldsymbol{Z}_j(w)$ rather than $\boldsymbol{H}_j(w)$), and 5, as well as assumptions from Athey et al. (2018) – 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 $\hat{V}_{\text{arb}} = \sum_{j=1}^{N} (\gamma_j^t - \gamma_j^c)^2 (Y_j - \boldsymbol{Z}_j \hat{\boldsymbol{\beta}}_{\text{lasso}})^2$.[19]

[19] In contrast to the variance in Theorem 1, the variance estimator $\hat{V}_{\text{arb}}$ in Theorem 2 is conditional on

Proof: See Appendix C.4.

4 Application: Snack Demand in the Laboratory

We test this approach in a setting of intrinsic interest to economists: estimating price elasticities for a collection of goods (Wright, 1928; Schultz, 1938; Stone, 1954). To ensure that we measure true price elasticities (ground truth) accurately, we employ a laboratory experiment in which we assess the demand for each good at two prices. Then we simulate treatment selection by subsampling to form a dataset containing a single demand observation for each good, and apply various estimators to the restricted dataset. In the interests of confronting our subjects with simple choices and evaluation tasks involving a reasonably large collection of familiar products, we settled on food items. In this section, each unit j is a food item; treatment $w \in \{0, 1\}$ represents a price of \$0.25 or \$0.75, respectively; $Y_j(w)$ represents aggregate demand at the price corresponding to w; and the outcome 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 items. In our setup, subjects make choices for different products independently; the method can accommodate substitution across products with slight modifications.

4.1 Experimental procedures and data

We estimated demand for snacks in a lab experiment with 365 subjects (181 males, 184 females).[20] Each subject was assigned to one of multiple treatments, described below. At the outset of each treatment session, subjects were told that the experiment would proceed in 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 of the experiment). Sessions took place in mid-afternoon, when subjects are typically hungry.

In the first stage of each session, subjects either made decisions that had a chance of being implemented (real choices $Y_j(w)$), or evaluations that were purely hypothetical ($\boldsymbol{H}_j(w)$). The hypothetical evaluations consisted of hypothetical choices for some subjects and subjective ratings for others. Real and hypothetical decisions pertained to snack food

hypothetical evaluations. Specifically, for a fixed sample size, the weights $(\gamma_j^t - \gamma_j^c)$ 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.

[20] We conducted the experiment at the Stanford Economic Research Laboratory (SERL); the first session was on November 15, 2010, and last on October 2, 2012. The protocol was reviewed and approved by Stanford University's IRB. Each subject received a participation fee between \$20 and \$30. We adjusted the fee upward when the response rate to our subject solicitation was low, and downward when it was high.

items offered at either \$0.25 or \$0.75. Subjective ratings pertained to the same collection of items, with price a factor in some but not all questions.

Each subject completed decision tasks or hypothetical evaluation tasks for $J = 189$ snack food items (at both prices, where applicable), with the stimuli (food items or item-price pairs) presented in random order.[21] Subjects were divided into multiple task-specific treatment groups, with each subject participating in a single treatment to avoid cross-contamination of responses across tasks. Most treatment groups consisted of roughly 30 subjects. For a complete catalog of the treatment groups along with sample sizes and a screenshot for a representative question, see Appendix D.1 and Figure A1.

4.1.1 Real choices

One group made real choices: they were informed that we would select one decision at random and implement it during the 30-minute waiting period. In observational data we would observe such demand at a single price, possibly set endogenously. Our experiment allows us to observe demand at both prices, which we use to establish ground truth. We then mimic observational data by making a supply assumption: we set a virtual price for each good and restrict the estimation sample to observations of demand at those prices.

A possible concern is that the low chance of implementing any given choice (one in 378 item-price pairs) renders it effectively hypothetical. Several checks strongly refute this concern. Despite the low implementation probabilities, subjects treated the real and hypothetical questions much differently. Average purchase frequencies are significantly higher for hypothetical choices than for these real choices (consistent with the general finding in the literature concerning hypothetical bias); 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. Additionally, the real choice frequencies do not change significantly when we increase the chance of implementation dramatically (to one in 5).[22] We are not surprised by the finding that participants in the "real choice" treatment viewed their choices as real. After all, they had as much at stake as someone making a single purchase decision (because they knew one choice would definitely be implemented), and their task was no more tedious when taken seriously.[23]

[21] The items belong to the following eight broad categories: candy (48 items), cookies and pastries (40 items), chips and crackers (24 items), produce and nuts (18 items), cereal (14 items), drinks (11 items), soups and noodles (11 items), and other (25 items).

[22] See Appendix D.3 for details.

[23] 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.

4.1.2 Hypothetical evaluations

Other participants provided various hypothetical evaluations, designed to span underlying motivations as well as factors that cause hypothetical choices to diverge from real ones.

Several groups made hypothetical choices. The literature on stated preferences explores a variety of protocols for eliciting such choices, and attempts to determine which is most accurate. However, it is not clear that any single approach dominates the others. Indeed, it seems likely that different protocols elicit different (and potentially complementary) information. Accordingly, we employed multiple protocols, each with a separate treatment group. One protocol mimicked the real choice treatment, except that no choice was implemented; we call this the "standard" protocol. A second protocol employed a "cheap talk" script (as in Cummings and Taylor, 1999) that encouraged subjects to take the hypothetical choices seriously,[24] 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 treatment groups provided subjective ratings. Depending on the group, subjects reported their anticipated degree of happiness with each potential purchase, the anticipated degree of social approval or disapproval for each potential purchase, how much they liked each item, evaluations of regret, measures of temptation, expected enjoyment (ignoring considerations of diet or health), perceptions of health benefits, impact of consumption on social image, and the perceived inclination to overstate or understate the likelihood of a purchase.

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

First we describe the true demand responses as implied by real purchases of each item at both prices (ground truth).

As expected, demand for a good falls as its price rises. On average, 28% of people elect to purchase a snack when the price is \$0.25; when the price is raised to \$0.75 only 20% do so ($p \leq 0.001$).[25] The demand for these products is relatively price inelastic, but the average response is sizable ($\tau = -7.5$ percentage points, standard error 0.4 percentage points).

Demand also varies among goods. The proportion who purchase varies from a low of 0 to a high of 60%, with a mean of 24%. We also see substantial variation in demand conditional

[24] 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).

[25] Throughout, when comparing two means, we use paired t-tests.

on price: the sample standard deviation is 11% with a price of $0.25 and 9% with a price of $0.75. While these statistics point to considerable heterogeneity in the attractiveness of the items, it is important to bear in mind that, given the size of the "real choice" treatment groups (30 subjects), some of that variation may reflect sampling uncertainty. There is also considerable variation across items in the responsiveness of demand to price: 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.[26]

4.1.4 Patterns of hypothetical choices

We also asked other participants whether they would hypothetically purchase each item at both prices. Not surprisingly, hypothetical choices exhibit substantial hypothetical bias: the average standard-protocol hypothetical demand (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 is more than twice that of real demand, a phenomenon we call hypothetical noise.[27] This noise does not appear to be due to hypothetical choices being more random than real choices (e.g., as might result if subjects took them less seriously). As we show in Appendix D.4, hypothetical noise is attributable in significant part to greater systematic variability of population hypothetical demand than of population real demand across choice problems. A possible explanation is that, when answering hypothetical questions, people naturally exaggerate the sensitivity of their choices to pertinent conditions; for example, as noted below, our data exhibit this pattern with respect to price variation.

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 choice 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 dots for item-price pairs with prices of $0.25, and blue dots for pairs with prices of $0.75. The relationship between hypothetical and real demand is systematic, and, helpfully for our purposes, stable between treatment

[26] Some of the variation in the measured price response across items is presumably attributable to sampling error, which differencing may either amplify or reduce, 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 4.4), 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.

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

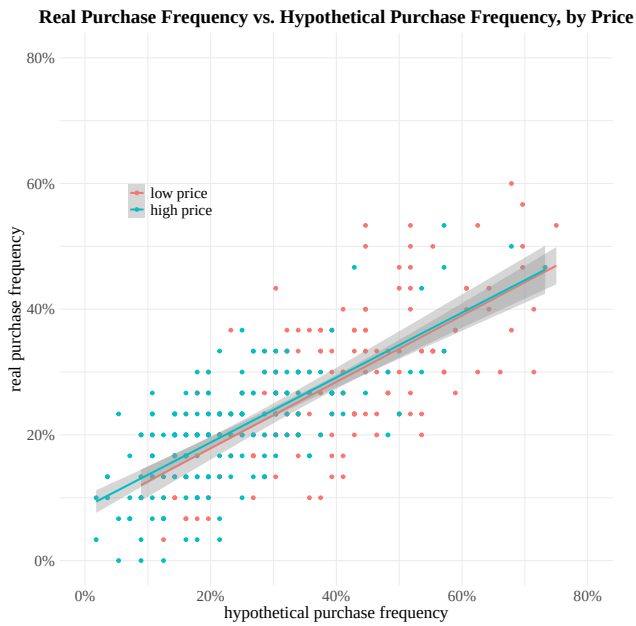


Figure 1: Real vs. Hypothetical Choices

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.

groups. To the extent we can identify the characteristics of choice problems that account for the greater variability of hypothetical choice frequencies for the population, we will be in a position to construct even better predictions of real choices.[28]

4.2 Estimation under endogenous treatment assignment

In this section, we mimic observational datasets in which each product is offered at a single price (treatment), and we observe the quantity sold (outcome). Prices vary across products, rather than for each individual product. To introduce endogenous treatment assignment, we first select a virtual price for each product that is correlated with potential outcomes, then drop the observation of the real choices at the other price.

4.2.1 Endogenous treatment

For the purpose of our analysis, we establish the price of each product based on respondents' hypothetical WTP for it, setting

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

[28] Visually, lowering the price (from blue to red) appears to shift the cloud to the northeast (higher hypothetical and real purchase frequencies) without disturbing the relationship between the variables.

for item j in simulation r. This procedure generates endogeneity in prices because WTP is strongly correlated with potential outcomes.[29] It simulates an environment in which sellers employ consumer surveys to assess the attractiveness of their products when choosing prices. (When deploying our method on the resulting data set, we do not allow it to access hypothetical WTPs. Our analysis therefore mirrors applications in which the analyst does not have access to the data that determine treatment assignment.) 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. 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.

4.2.2 Results

We first compare the accuracy of simple (univariate) versions of the estimators proposed in this paper to that of some simple standard estimators discussed in the literature. Table 1 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 a price increase from \$0.25 to \$0.75 affects the fraction of subjects actually buying the average snack by -0.075.

Simple standard estimators exhibit substantial bias. Taking the difference in means (mean of treated minus mean of control, Column (2)) yields an estimated effect of -0.025. As in real applications, the endogeneity of prices and quantities leads this estimator to understate price sensitivity substantially.

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.

The literature on stated preferences considers variants of the hypothetical choice protocol that are intended to "fix" this hypothetical choice bias. While we find that some of the alternative protocols reduce the overall degree of hypothetical bias compared with the standard protocol, it appears that they generally do so in our experiment 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

[29] Appendix Figure A2 plots a snack's actual purchase frequencies 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 similar conclusions.

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.[30] 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 preferring this estimate to less accurate ones that employ different dichotomization thresholds. 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 between high and low prices). 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), which suggests that posing vicarious choice questions does not improve the informational content of the response. The estimated treatment effect from the vicarious question may be smaller simply because there is a greater likelihood that people respond randomly rather than informatively, which attenuates the difference between the means. It is particularly striking that 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 the hypothetical choice measures. However, all of these hypothetical responses are clearly correlated with real choices, and thus may make useful predictors.

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 1 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

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

Table 1: 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.

that more comprehensively span motivations. Table 2 explores this possibility. For convenience, Column (1) reproduces the true average treatment effect. The next two columns investigate whether it is possible to obtain more accurate estimates of treatment effects by controlling for more 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)).[31] For our method, we exhibit results based on several reasonable 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 3.2. 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 D.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

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

Table 2: 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 ($\times 2$)	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.

(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 1 (-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.

4.3 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, a feature that renders conventional observational estimation infeasible. In such environments, unconfoundedness (Assumption 1) is satisfied trivially, but estimation relies on the accuracy with which hypothetical responses can 'extrapolate' into the unseen setting. Theoretically, extrapolation is accurate when the mapping from predictors to outcomes is invariant (Assumption 3), as long as either the distributions of evaluations for the hypothetical treatment states are overlapping (Assumption 5) or the relationship is linear (Assumption 4).

4.3.1 Results

Table 3 shows that even if the 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

Table 3: 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.

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).[32] 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 evaluations overlap assumption. Columns (4)–(8) employ specifications analogous to those in Table 2, 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 further from the truth.

4.3.2 Discussion

When predicting the outcome in an unseen treatment state, our method projects from the space of treatment, where variation is absent, into motivation space, where characteris-

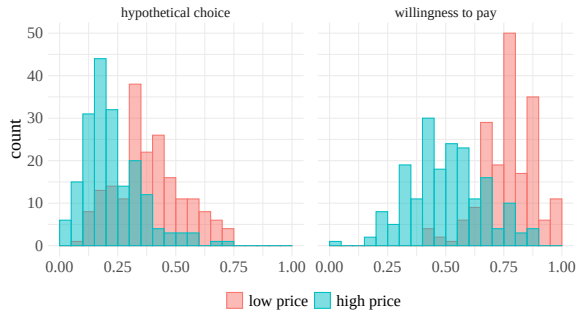
[32] Estimates for the other univariate specifications in Table 1 are in the Online Appendix.

tics 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. Overlap can (and should) be diagnosed directly in applications, even without observation of ground truth. Part (a) assesses overlap using histograms, which depict the distributions of evaluations for the high price in blue and for the low price in red. 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 outcome 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 less than half of 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 much greater extrapolation. Hence, for the WTP choice, we can predict more confidently from high price to low price than in the opposite direction.

Part (b) uses our ground truth 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 observe the red dots, from which we might infer the red 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 at that point. We can discover this in our experiment (for which we actually have real choices at both prices) by inspecting the high-price data (the blue 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 4). When using our method for unseen counterfactuals, care should be taken to inspect and justify overlap. While one can also check linearity, our example strikes a cautionary note: a relationship can be linear only within the observed (overlapping) range of variation. Hypothetical data are more likely to satisfy overlap when the variation from the treatment is small relative to that arising from other factors. Ideally, the variation for each type of hypothetical evaluation in the observed treatment state (for example, the level of temptation and degree of social approbation) should span the variation for the unobserved treatment state.



(a) Overlap in Hypotheticals

(b) Relationship between Outcome and Hypothetical

Figure 2: Overlap between Hypothetical Evaluations

4.4 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.

4.4.1 Metrics

We report four measures of the degree to which the estimated heterogeneity in treatment effects, $\hat{\tau}_j$ for unit j, captures the heterogeneity in actual effects, τ_j:

- R^2 for a regression of the true treatment effect $\tau_j = Y_j(1) - Y_j(0)$ on the predicted treatment effect $\hat{\tau}_j = \hat{Y}_j(1) - \hat{Y}_j(0)$: This statistic measures the fraction of the variation in true treatment effects that the estimated treatment effects capture.

- *Mean squared error* ($\text{mse} := \text{mean}((\tau_j - \hat{\tau}_j)^2)$)*:* This statistic encompasses both the overall accuracy and the precision of unit-level estimates.

- *Calibration coefficient:* This is the slope coefficient in a regression of the true treatment effect τ_j on the predicted treatment effect $\hat{\tau}_j$. The ideal value of this statistic is 1,

indicating that the expectation of the actual treatment effect varies unit for unit with the predicted treatment effect.[33]

- *Simulated profit:* We simulate a producer who estimates the (heterogeneous) price sensitivity in order to set optimal prices. Let true profit for snack j be $\pi_j(w) = (w \cdot 0.75 + (1-w) \cdot 0.25 - c)Y_j(w)$ for $w \in \{0, 1\}$. 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.[34] The producer observes demand for snack j at a single price W_j, and predicts demand at the other price: $\hat{Y}_j(w) = Y_j(w) + \hat{\tau}_j \cdot (1_{\{w>W_j\}} - 1_{\{w<W_j\}})$. The producer sets the price to maximize predicted demand: $w_j^* = \arg\max_w (w \cdot 0.75 + (1-w) \cdot 0.25 - c) \cdot \hat{Y}_j(w)$. We report the gain in average profit, $\bar{\pi}(\mathbf{w}^*) = \frac{1}{J} \sum_j \left[(w_j^* - c)Y_j(w_j^*)\right]$, over the average profit derived from setting the prices at random, which we express as a fraction of the maximum possible gain achieved by optimal pricing.[35] The firm achieves optimal profits when $\hat{\tau}_j = \tau_j$ for all j. Imperfect estimates of treatment effects cause the firm to deviate from optimal prices and result in lower profits.

To provide a consistent benchmark, in all cases we compare estimates to the true treatment effect for unit j, τ_j, rather than a treatment effect averaged over units similar to j (that is, the conditional average treatment effect $E(\tau_j|Z = z)$ for some set of covariates Z). Hypothetical evaluations can both improve the estimation of average treatment effects that condition on a given set of covariates, and also substantially enrich conditioning sets.

4.4.2 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.

[33] 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 $\frac{\text{var}(\tau_j)}{\text{var}(\tau_j)+\text{var}(\epsilon_j)} \leq 1$.

[34] 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.

[35] We focus on the simplest plausible rule; other pricing policies may perform better according to some metrics because we estimate optimal prices with variance.

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 estimator 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.

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.[36] 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 3 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 potentially capture substantially more unit-level heterogeneity than conventional methods, we next ask whether it uses the information contained in physical characteristics less, equally, or more effectively. 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 selection is endogenous, our method still performs well in terms of recovering heterogeneous effects according to all four metrics. For row 8 of Figure 3, we use the

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

same prediction model as in row 6, but we apply our method to simulated draws based on the endogenous assignment rule described Section 4.2.1. 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 3. Our method noticeably outperforms feasible conventional approaches that condition on physical characteristics even when we handicap it (and not the alternative methods) by introducing endogenous treatment selection.

4.4.3 Discussion

Our method provides a way to recover heterogeneous treatment effects even without randomized experiments. Quasi-experimental methods identify causal effects only for small subsets of the data that are less confounded (e.g., around a discontinuity, or among compliers in IV). As a result, they are ill-suited for measuring heterogeneity other than within special subsets for which variation is roughly exogenous. For that reason, most analyses of heterogeneous treatment effects rely in practice on randomized experiments that alter treatment states across the board (though these can still only measure effects for units that comply). But such sweeping interventions are possible only in some settings. 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.

Our method also may recover finer, more robust heterogeneity. With standard methods and data, estimating detailed conditional average treatment effects is fundamentally difficult because potential nonlinearities require local estimation, reducing the effective sample size. Estimation methods that separately model outcomes under both treatment states further reduce the available data by half because they divide it into treated and control samples. In contrast, our method infers the mapping from hypothetical evaluations to outcomes using data on all units. When hypothetical choice biases are systematic, relatively simple functional forms can offer reasonable *global* approximations because one measures real and hypothetical choices on the same scale. Moreover, when standard methods must make do with observing each unit j in a single treatment state, our method estimates the treatment effect for unit j based on the unit-specific hypothetical responses in *both* states. These hypothetical evaluations often contain information not readily captured by standard covariates. Hence, our method can yield estimates of conditional average treatment effects (CATE) substantially closer to the true *unit-level* effects than those that condition only on standard covariates.

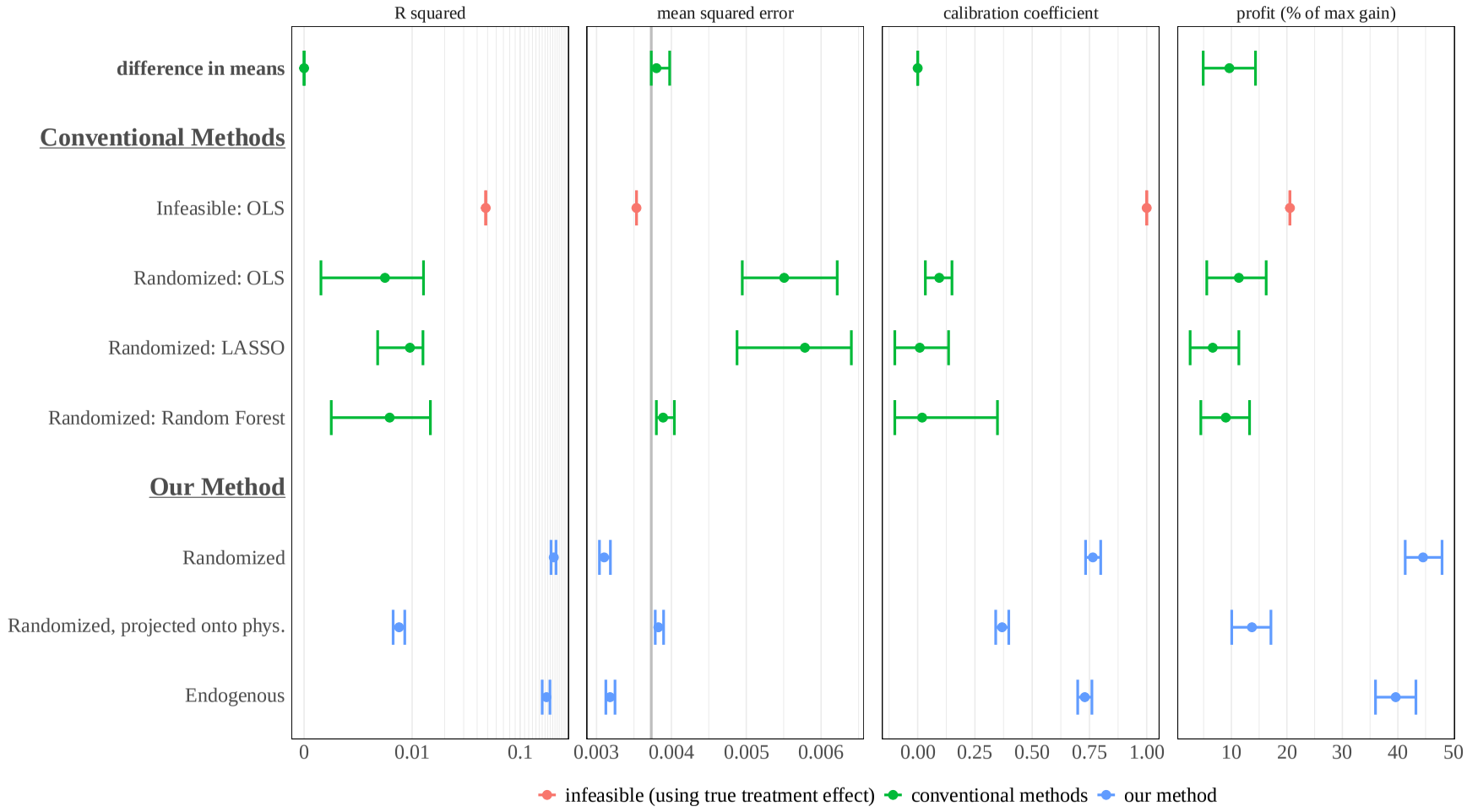


Figure 3: Treatment Effect Heterogeneity

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. R-squared is from a regression of the true treatment effect τ_j on the predicted treatment effect $\hat{\tau}_j$; calibration is the coefficient on $\hat{\tau}_j$ in the same regression; mean squared error is the average squared distance between τ_j and $\hat{\tau}_j$; profit is for the simulated price-setting exercise described in the text. 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. For profit, we show the gain over random pricing, expressed as a fraction of the maximum achievable gain (with accurate knowledge of treatment effects).

4.5 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 unbalanced, while our method maintains good performance even if few of the observations are treated (or none, as in Section 4.3). 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 the implied 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 2, 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 3.[37] Figure 4 plots the resulting statistics as functions of the fraction of snacks observed at the high price.

The first panel of Figure 4 shows that the standard deviations of the estimators that employ our method are substantially smaller than those of the conventional estimators. While these standard deviations hold fixed the sample of snacks, the standard error formulas also reflect the additional variation associated with sampling snacks (independently) from some super-population.[38] The (median) standard error of the difference-in-means is more than twice that of the univariate hypothetical choice estimator when half of the sample is treated, the most favorable balance for conventional estimators.[39] 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.

The comparison becomes even more favorable to our method for unbalanced designs.

[37] Figures including all specifications of our method from Tables 1 and 2 are in the Online Appendix.

[38] 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 numerically obtain similar patterns for estimated standard errors also for the other estimators.

[39] Figure 4 presents simulated standard deviations; to see panels for estimated standard errors and coverage, see Appendix Figure A4.

The standard deviation of the difference-in-means is U-shaped in the fraction of treated observations: because the means of the treatment and control groups are estimated from separate subsamples, the smaller subsample dominates the variance. Our low-dimensional estimator, in contrast, pools all observations in the first step to estimate the relationship between outcome and hypothetical evaluations. In the second step, we use the hypothetical evaluations of both treatment states for each snack, so again there is no direct dependence on the fraction of observations in the treated state. Thus, the precision of our low-dimensional estimator is largely independent of this fraction.

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.

In this application, a smaller standard deviation comes at the cost of a small bias (Figure 4 second panel), but our estimators remain preferable to standard alternatives in terms of 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 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. 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. The high-dimensional version of our method reduces this 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. 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.

5 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 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 a microfinance website. We evaluate the method's accuracy through comparisons with a ground truth

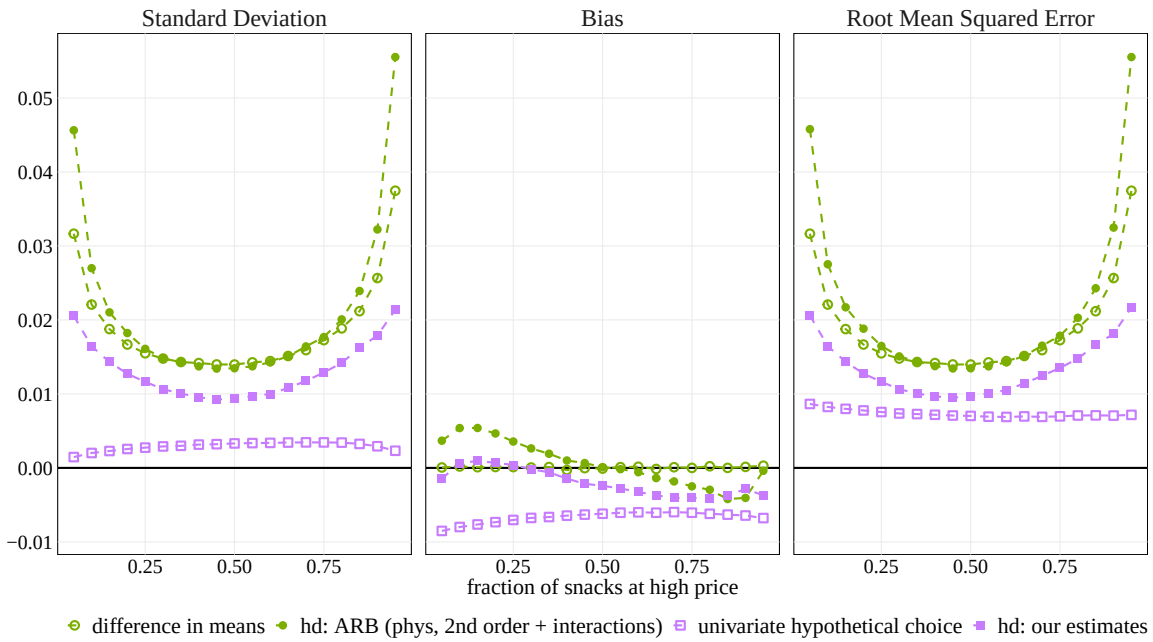


Figure 4: Performance of Estimators by Fraction Treated

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 4.3), and the standard deviation (across realizations of the assignment distribution) is mechanically zero because there is only one possible assignment.

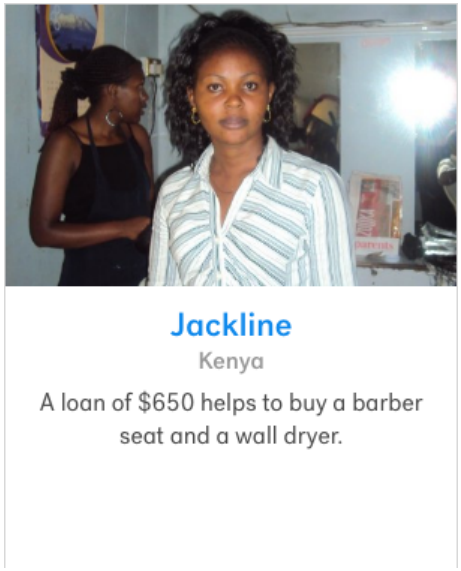
estimate based on an experiment in which we introduce randomly assigned matches.[40]

Our analysis focuses on a large microfinance crowdsourcing website, which displays profiles of potential borrowers and allows website visitors to contribute to their loans. See Figure 5 for examples of a standard and matched loan. The assignment of a match involves a complex process. In particular, the website cultivates sponsors who provide funds for matching loans, and who 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). If a loan profile meets an active sponsor's criteria, the website displays it with a matching indicator.

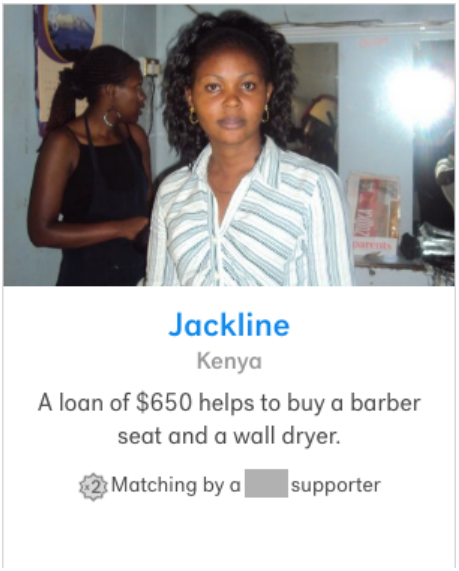
Correlations between the preferences of sponsors and contributors render the treatment endogenous. Some of these correlations can, in principle, be controlled for if all possible matching criteria are observed. Our method may nevertheless have the advantage of using reasonably parsimonious specifications, whereas controlling for all possible combinations (interactions) of loan characteristics requires an extremely high-dimensional model.[41] More fundamentally, however, some of the endogeneity in this application may be difficult or impossible to address through the inclusion of controls. For example, complications arise

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

[41] Similarly, using hypothetical evaluations in our method can be substantially more parsimonious than controlling flexibly for image or text data that affect both treatment assignments and outcomes.



(a) Unmatched profile



(b) Matched profile

Figure 5: Loan Profiles with Matching Indicator

if sponsors decide which types of profiles to match based in part on the attractiveness of postings within particular categories at the time of the matching decision.

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.[42] The treatment effect of primary interest is the average impact of matching on fundraising.

5.1 Data

In this section, we describe the observational data, experimental data, and survey data on hypothetical responses used in our analysis.

5.1.1 Observational data

We observe 11,668 loan profiles for borrowers seeking $1,000 or less posted to the website between October 14, 2019, and November 3, 2019, that were not matched as part of our experiment. Of these loan profiles, we keep 9623 profiles (82%) that we can classify as matched (because they were matched for at least 90% of the first 24 hours after their initial

[42]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.

posting) or unmatched (because they were matched for 3% or less of this time).[43] In this sample, 623 (6.5%) of the profiles were classified as matched. For each of these profiles, our data include descriptive characteristics, when it was matched, and how quickly it raised funds.

5.1.2 Ground truth experiment

In this application, the endogeneity of the treatment makes it difficult to obtain reliable estimates of treatment effects from the available observational data. Plausible candidates for instrumental variables are difficult to identify, and other estimation strategies are not promising.[44] For this reason, 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%).[45] 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. Consequently, some loans in the control group were match-eligible, and some in the treatment group would have been match-eligible without our intervention. 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 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 the population of unmatched loans in the observational sample, and the local average treatment effect (LATE) corresponds to the average treatment effect on the control (ATC). Loans that are matched in our experiment irrespective of our intervention are always-takers;

[43] We drop the 18% of profiles that were matched for part but not all of their first 24 hours to create a binary treatment indicator. The estimated effect of matching based on the observational data is very similar if we use all profiles and specifications that are linear in the share of the first 24 hours that each profile was matched.

[44] In principle, one might be able to exploit discontinuities arising from the depletion and replenishment of matching funds, but in practice we do not have access to information on sponsors' balances or their matching criteria, which makes the points of discontinuity difficult to identify. Even if such data were available, the procedure would recover the LATE for a highly selected population, inasumch as contributions are more likely to exhaust matching funds targeting more appealing loan profiles. Alternatively, because sponsor's criteria must reference a known collection of observable characteristics, one might simply regress the outcome on the treatment, controlling for these factors. In practice, the list of characteristics is long, and sponsors can specify multiple conditions (e.g., African women seeking loans to finance agricultural projects). Controlling for all conceivable categories of loans (i.e., all permutations of factors) is infeasible.

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

they correspond to matched loans in the observational sample. The effect on always-takers corresponds to the average treatment effect on the treated (ATT). 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).

5.1.3 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 to allow more precise estimation of the ATT. 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 projecting 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 projected 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 E.1.

5.2 Local Average Treatment Effects

Table 4 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. The resulting coefficient for the matching dummy is 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, even knowing the structure of the process that renders matching provisions

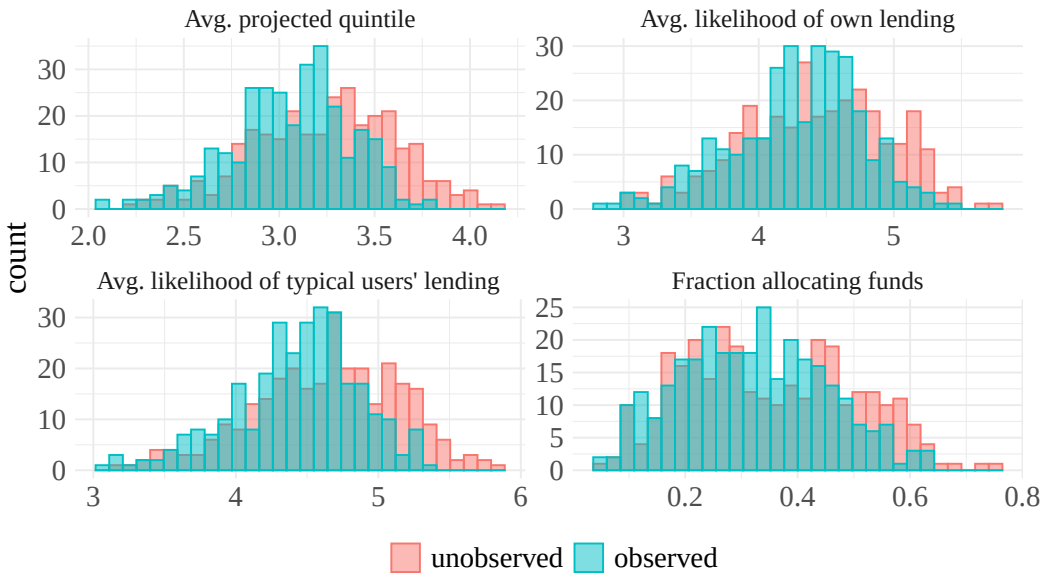


Figure 6: Overlap in hypothetical evaluations for loan × treatment states that are observed (blue) vs. unobserved (red) in the data.

endogenous is challenging, 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 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 experimental results both economically and statistically. Table 4 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)), frequencies of each possible hypothetical response (i.e., distributional detail) (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

Table 4: 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.

truth.

5.3 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 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 5. The first row reproduces selected estimates of the LATE (also the ATC) from Table 4, including the IV estimate, as well as two measures obtained through our method (corresonding to the low and high dimensional specifications

Table 5: 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 4, columns (1), (5), and (9) (as indicated in the column headings). Standard errors in parenthesis are based on the bootstrap.

in, respectively, columns (5) and (9) of Table 4). 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 widely among compliance groups. The second row shows that our estimates of the average treatment effect on the treated (ATT) is less than half as large as the LATE/ATC for both specifications. Loans that are matched in practice do not benefit as much from the match, presumably because they are sufficiently attractive along other dimensions to achieve high fundraising velocity 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. By way of analogy to our analysis of optimal snack pricing in Section 4.4, 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.

6 Extensions

6.1 Heterogeneity in ability to predict real choices among survey respondents

In some applications, the survey respondents answering hypothetical questions may differ noticeably from the people whose choices determine the real outcomes. For example, in our microfinance application, visitors to the website determine the outcome of interest, but we obtain hypothetical evaluations by drawing a sample of respondents from Amazon

Mechanical Turk, fewer than 25% of whom report having visited the website.[46] One possibility is to screen survey respondents based on understanding questions, attention checks, and their reported characteristics (such as interest in microfinance). Here we describe a data-driven alternative, which identifies and relies upon the subset of survey respondents who demonstrate the greatest ability to predict real decisions.

We propose a variant of our method in which we filter responses based on a measure of *latent response quality* r_{kj} for each respondent k's evaluation of setting j. We define r_{kj} as the correlation of k's evaluations with outcomes *for other settings* $j' \neq j$. When implementing our estimation method, we only include observations (k, j) for which this latent response quality exceeds some threshold: $r_{kj} \geq r$.[47] Because this procedure reduces the number of evaluations per setting, it raises small sample concerns: we observe a small random sample of $\boldsymbol{H}_{kj}(w)$ rather than the population measure $\boldsymbol{H}_j(w)$. Under our assumptions, $\mathbb{E}(\boldsymbol{H}_{kj}(w)) = \boldsymbol{H}_j(w)$, so this consideration implies that we measure $\boldsymbol{H}_j(w)$ with error. We replace step 1 of our method with instrumental variables on split samples (for instance, Fuller, 1987) to remove the attenuation bias.[48]

We illustrate this procedure by applying it to the problem of measuring the effects of the microfinance website's matching provisions. We filter based on how strongly respondents' "quintile projection" correlates with actual fundraising velocities, based on loan profiles displayed in their actual treatment states.[49] 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.

A possible criterion for selecting a threshold r^* is to minimize the mean squared error out-of-sample.[50] The selected correlation threshold ($r^* = 0.16$) is shown as a vertical line in Figure 7. For that threshold, the procedure yields an estimated treatment effect of 1.6.

Because similar thresholds mechanically yield similar average evaluations, the optimal threshold is imprecisely estimated in finite samples. As a more robust option, we propose using a residual balancing approach similar to the one deployed for our main high-dimensional

[46] 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 lab experiment 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 lab experiment in Appendix Figure A5 separately with all snacks observed at the high price and all snacks observed at the low price.

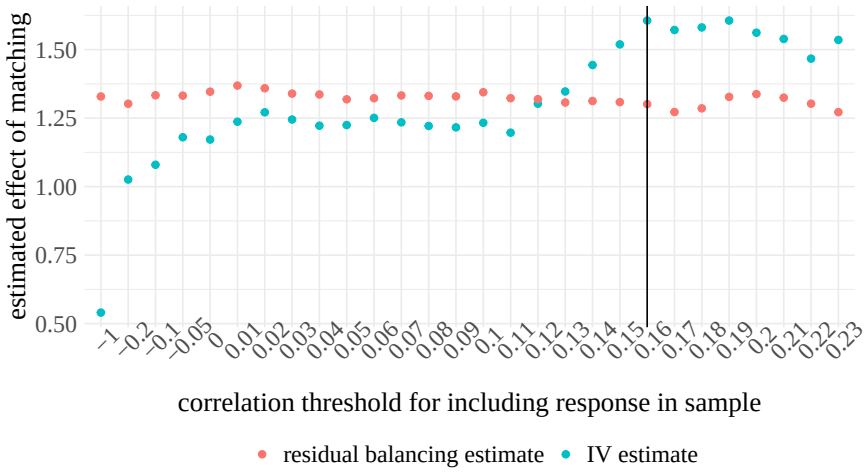
[47] This leave-one-out correlation r_{kj} avoids overfitting by omitting any direct information on the predictive accuracy of k's evaluation for the j-th setting.

[48] 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. We report the median across 11 such random sample splits.

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

[50] See Appendix E.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



estimator. In particular, imagine estimating a penalized regression including all variables measured at a large (but finite) collection of thresholds spanning the range of possibilities. Selection of the optimal threshold is similar to subset selection with subset size set to unity. Residual balancing adds a correction to the instrumental variables estimates so that even if the chosen correlation threshold is not (asymptotically) optimal, bias would be small in any linear model. Figure 7 shows the residual balancing estimates (the red dots) one would obtain if each of the thresholds was selected (because a different criterion was used for selection). As shown in the figure, the residual balancing estimates are largely unaffected by the selected threshold, because they balance all thresholds simultaneously. All of the resulting estimates are between 1.27 and 1.37. The estimate for the threshold selected for our IV procedure ($r^* = 0.16$) is 1.30.

6.2 Treatment as choice

Estimation of causal effects is commonly confounded because the treatment is correlated with potential outcomes. One can make progress by modeling the process by which either the outcome or treatment is determined (or both), as is apparent from the literature on doubly robust estimation of treatment effects (Robins and Rotnitzky, 1995; Chernozhukov et al., 2018). This paper focuses on modeling the outcome, and we develop assumptions and estimators for applications where the outcomes result from human choice ("outcome as choice").

An alternative approach is to model the process by which treatment is determined ("treatment as choice"). Modeling the treatment propensity score, rather than the outcome, appears most natural if we can elicit hypothetical evaluations from the entity choosing the

treatment. In "treatment as choice" applications, certain kinds of evaluations, such as stated treatment probabilities, may resemble (monotone transformations of) the latent single index in a Roy model of selection into treatment. That could make it possible to estimate the effects of policies that shift the threshold for treatment (Briggs et al., 2020).

Eliciting hypothetical evaluations or treatment probabilities from the entity making the treatment choice presents several challenges. First, the party or parties controlling the treatment (e.g., pricing authorities for the product demand application, benefits managers for the 401(k) application) may be difficult to survey. Second, obtaining hypothetical responses from the same individuals who make the treatment choices can introduce confounds, because people tend to resolve uncertainty about choices over time. If a person is asked to evaluate the decision hypothetically before they make a decision about the treatment, then the evaluations may not include all the information they will have when they make the actual choice (potentially violating unconfoundedness). Additionally, being asked can distort choices (Zwane et al., 2011). Alternately, if the person is asked for hypothetical evaluations after they have already made a decision about the treatment, they may distort their hypothetical evaluations for the sake of consistency with that choice. That tendency can create a violation of unconfoundedness (in that it induces a bias in the response), or overlap (if no one who chose a particular treatment says they may have chosen the other option under a different scenario). More work is needed to identify the characteristics of applications for which this strategy yields credible estimates.

One can alternately model the processes determining both outcome and treatment. We outline a version of that approach, one that employs a doubly robust estimator and combines propensity score and outcome modeling, in Appendix C.3. However, the estimator does not retain all the attractive features that characterize the low- and high-dimensional approaches to outcome modeling, despite incorporating the same structural assumptions. Hence, when using hypothetical evaluations, there may be benefits to modeling outcomes only in "outcome as choice" applications.

7 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 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 allow the analyst to recover average treatment effects even in settings where standard methods are inapplicable due to the absence of plausible instruments or helpful discontinuities. Second,

one can apply it even in cases for which there is no observed variation in the treatment. Third, it yields more comprehensive measures of heterogeneous treatment effects than standard approaches, in that it allows the analyst to 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 experiment, and through a field application involving the effects of matching loan provisions offered on a large microlending platform.

We do not claim that the approach offers a panacea. On the contrary, our objective has been to articulate the conditions under which such an approach should work in order to facilitate judgments concerning its suitability for any given application. 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 offering a low-cost method for exploring 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 assumes portability within a class of decision problems, but this is a 'local' assumption. 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

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 Dis-

crepancy 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, 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.

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.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz. 2004. "The short-and long-term effects of measuring intent to repurchase." *Journal of Consumer Research* 31 (3): 566–572, Publisher: The University of Chicago Press.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz. 2005. "Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research." *Journal of Marketing* 69 (2): 1–14. 10.1509/jmkg.69.2.1.60755, Publisher: SAGE Publications Inc.

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.

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.

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.

Engen, Eric M., William G. Gale, and John Karl Scholz. 1996. "The illusory effects of saving incentives on saving." *Journal of economic perspectives* 10 (4): 113–138.

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.

Kraut, Robert E., and John B. McConahay. 1973. "How Being Interviewed Affects Voting: An Experiment." *Public Opinion Quarterly* 37 (3): 398–406. 10.1086/268101.

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.

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.

Maslow, Abraham Harold. 1943. "A theory of human motivation.." *Psychological review* 50 (4): 370, Publisher: American Psychological 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.

Poterba, James M., Steven F. Venti, and David A. Wise. 1996. "How retirement saving programs increase saving." *Journal of Economic Perspectives* 10 (4): 91–112.

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." *Econometrica* 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.

Zwane, Alix Peterson et al. 2011. "Being surveyed can change later behavior and related parameter estimates." *Proceedings of the National Academy of Sciences* 108 (5): 1821–1826. 10.1073/pnas.1000776108.

Appendices

A Appendix Figures

Figure A1: Demand Experiment: A typical choice task

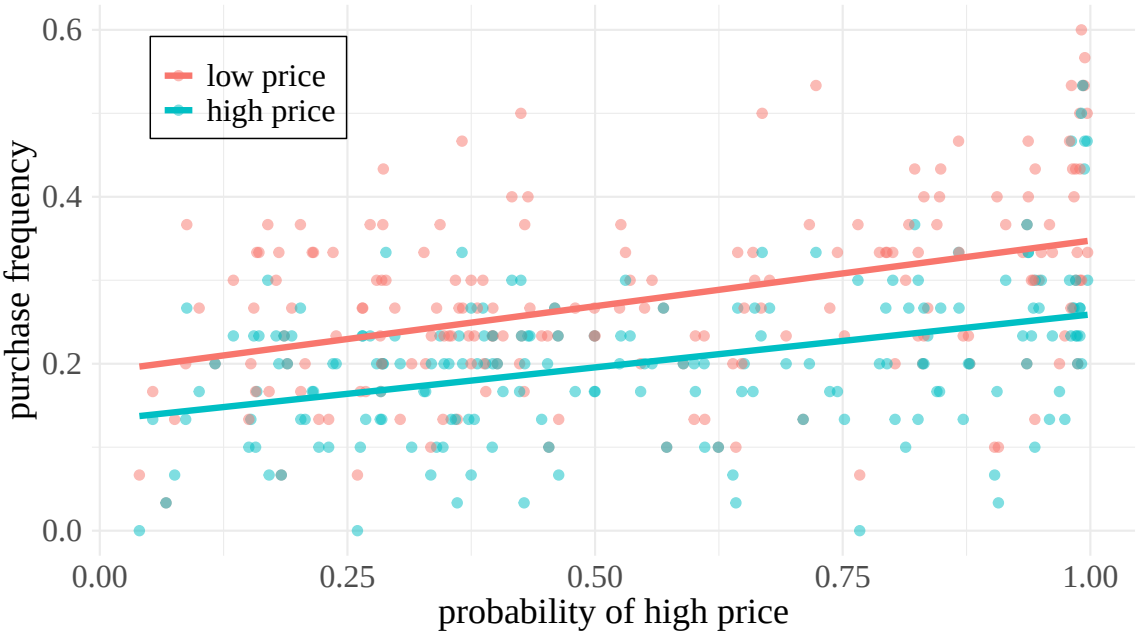


Figure A2: Simulation setup

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.

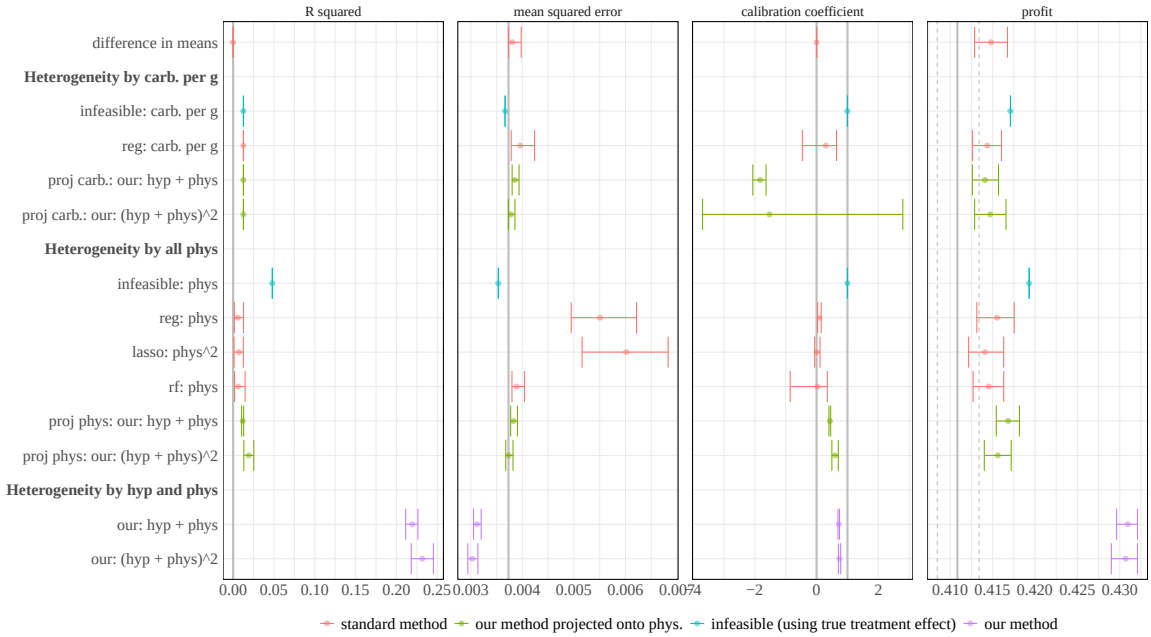


Figure A3: Summary statistics describing how well different estimators describe heterogeneity in treatment effects, with high-dimensional methods. Points show the median statistics across 1,001 simulated samples, and error bars indicate the interquartile range in the simulations.

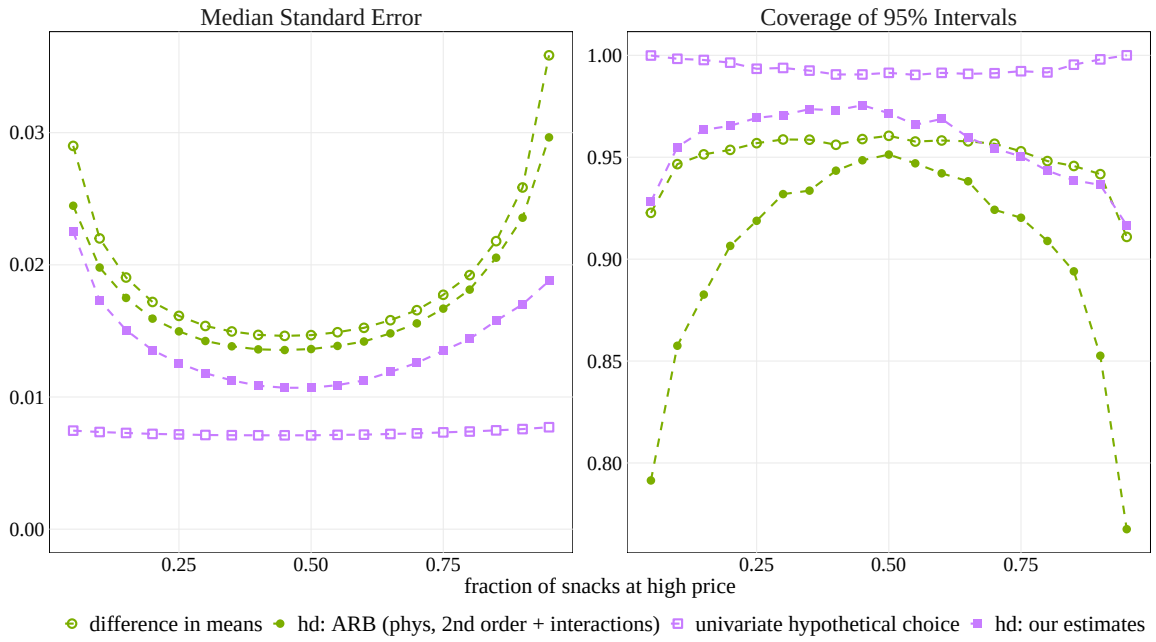
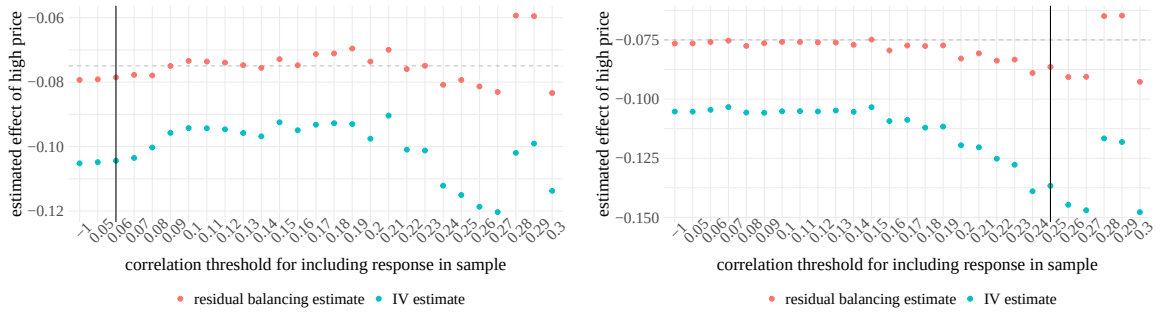


Figure A4: Performance of Estimators by Fraction Treated: median standard error and coverage (across samples differing by treatment assignment).



(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 line indicates the threshold selected by mean square error fit of the step 1 regression. The dashed 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);[51] 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).[52]

It is well-established that answers to standard hypothetical questions are systematically biased.[53] Two classes of solutions have been examined. One attempts to "fix" the hypothetical question.[54] Our approach is more closely related to a second class of solutions involving ex post statistical calibration.[55] 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

[51] 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).

[52] 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).

[53] The bias typically favors overstatement of willingness-to-pay and alternatives that are viewed as more "virtuous." See, for example, Cummings et al. (1995), Johannesson et al. (1998), List and Gallet (2001), Little and Berrens (2004), Murphy et al. (2005), and Blumenschein et al. (2008). 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.

[54] 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).

[55] 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).

across decision problems. Consequently, they are not useful for predicting choice distributions in decision problems that have not yet been observed.[56] On the contrary, List and Shogren (1998; 2002) emphasize that hypothetical bias is context-specific, so that individual-level calibration does not reliably transfer from one setting to another.[57] 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.[58] 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,[59] 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

[56] 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 also available.

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

[58] While Blackburn et al. (1994) do not reject the hypothesis of no contamination, their test is limited to a single setting and its power is unclear. Moreover, marketing studies have found, on the contrary, that stated intentions influence subsequent choices (see, e.g., Chandon et al. (2004; 2005)). Similarly, voter surveys have been shown to affect turnout (see, e.g., Kraut and McConahay (1973)).

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

reactions predict choices.[60] 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 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.[61] 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 Proofs and additional theoretical results

C.1 Proof of Theorem 1

The data are a random sample of independent observations $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1))_{j=1}^J$ where $Y_j \in \mathbb{R}$, $W_j \in \{0,1\}$, and $\boldsymbol{H}_j(1), \boldsymbol{H}_j(0) \in \mathbb{R}^q$ are row vectors. Define $\boldsymbol{H}_j = \boldsymbol{H}_j(W_j)$. The estimator proceeds in two steps: first, regress outcomes Y_j on hypothetical evalations $\boldsymbol{H}_j$. Second, take the estimated coefficients on $\boldsymbol{H}_j$, say $\hat{\boldsymbol{\beta}}$, and calculate $\hat{\tau} = \frac{1}{J}\sum_{j=1}^J (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\hat{\boldsymbol{\beta}}$.

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

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

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

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

By Assumptions 2, 3, and 4,

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

where $\tau^* = \mathbb{E}(Y_j(1) - Y_j(0))$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}_{0,0} = \boldsymbol{\beta}_{1,1}$ with $\boldsymbol{\beta}_{w,w}$ as specified in Assumption 4. The equality between $\boldsymbol{\beta}_{0,0}$ and $\boldsymbol{\beta}_{1,1}$ hold by Assumptions 2 and 3. Assumption 2 further implies that $\boldsymbol{\beta}_{0,1} = \boldsymbol{\beta}_{1,0} = 0$. By Assumptions 1, 2, 3, and 4 and random sampling,

$$\mathbb{E}(\boldsymbol{m}(Y_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{H}_j, \tau^*, \boldsymbol{\beta}^*) = \mathbf{0}_{q\times 1}.$$

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

Define

$$\begin{aligned}\boldsymbol{\Gamma} &= \mathbb{E}\Big(\frac{\partial \psi(Y_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{H}_j, \tau^*, \boldsymbol{\beta}^*)}{\partial(\tau, \beta)}\Big) \\ &= \mathbb{E}\left(\begin{bmatrix} 1 & -(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0)) \\ \mathbf{0}_{q\times 1} & -\boldsymbol{H}_j'\boldsymbol{H}_j \end{bmatrix}\right)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Psi} &= \mathbb{E}(\boldsymbol{\psi}\boldsymbol{\psi}') = \mathbb{E}\left(\begin{bmatrix} g^2 & g\boldsymbol{m}' \\ g\boldsymbol{m} & \boldsymbol{m}\boldsymbol{m}' \end{bmatrix}\right) \\ &= \mathbb{E}\left(\begin{bmatrix} (\tau^* - (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\boldsymbol{\beta}^*)^2 & \boldsymbol{H}_j(\tau^* - (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\boldsymbol{\beta}^*)(Y_j - \boldsymbol{H}_j\boldsymbol{\beta}^*) \\ \boldsymbol{H}_j'(\tau^* - (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\boldsymbol{\beta}^*)(Y_j - \boldsymbol{H}_j\boldsymbol{\beta}^*) & \boldsymbol{H}_j'\boldsymbol{H}_j(Y_j - \boldsymbol{H}_j\boldsymbol{\beta}^*)^2 \end{bmatrix}\right)\end{aligned}$$

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

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

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

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

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

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

$$\begin{aligned} V_\tau = \; & \mathbb{E}\Big((\tau^* - (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\boldsymbol{\beta}^*)^2\Big) \\ & + \mathbb{E}\Big(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0)\Big) \boldsymbol{V}^{\text{ols}} \mathbb{E}\Big(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0)\Big)' \\ & - 2\mathbb{E}\Big(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0)\Big) \mathbb{E}\Big(\boldsymbol{H}_j'\boldsymbol{H}_j\Big)^{-1} \mathbb{E}\Big(\boldsymbol{H}_j'(\tau^* - (\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\boldsymbol{\beta}^*)(Y_j - \boldsymbol{H}_j\boldsymbol{\beta}^*)\Big) \end{aligned}$$

C.2 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, 3, and 5.*

Proof: $\mathbb{E}(Y_j(1) - Y_j(0)) = \mathbb{E}\big(\mathbb{E}(Y_j(1) - Y_j(0) \mid \boldsymbol{H}_j(1), \boldsymbol{H}_j(0))\big)$ *by the law of iterated expectations. The next steps hold for $w \in \{0, 1\}$. By Assumption 2, $\mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(1), \boldsymbol{H}_j(0)) = \mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(w))$. $\mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(w) = \boldsymbol{h}) = \mathbb{E}(Y_j \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, W_j = w)$ by unconfoundedness Assumption 1. $\mathbb{E}(Y_j \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, W_j = w) = \mathbb{E}(Y_j \mid \boldsymbol{H}_j(W_j) = \boldsymbol{h})$ is identified by Assumptions 3 and 5 for all relevant levels of $\boldsymbol{h}$.*

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.

C.3 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 3, it is easy to verify that the following moment condition satisfies the Neyman orthogonality condition:

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

where μ is the relationship between outcome and hypothetical evaluations of the realized treatment state, and $e_w(\boldsymbol{h}) = \Pr(W_j = w | \boldsymbol{H}_j(w) = \boldsymbol{h})$ 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. 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. Interestingly, this doubly robust moment, despite using the same structural Assumptions 2 and 3, also requires a more standard overlap assumption bounding conditional treatment probabilities away from 0 and 1. Consequently, it cannot be used to estimate the effect of an unseen treatment. 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.

C.4 Proof of Theorem 2

The result follows from Lemma 2 of Athey et al. (2018) by noting that our unconfoundedness Assumption 1 replaces their Assumption 1, our Assumptions 2, 3, and 4 jointly replace their Assumption 2, and our overlap Assumption 5 replaces their Assumption 6. 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 $\boldsymbol{\gamma}^t$ and $\boldsymbol{\gamma}^c$ 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 $\boldsymbol{\gamma}^t - \boldsymbol{\gamma}^c$ takes the role of $\boldsymbol{\gamma}$ in the "Proof of Corollary 6" in their on-line appendix.

D Snack Demand Application

D.1 Treatment groups

Treatment 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.

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

Treatment M (35 subjects): Subjects considered the same choice problems as in treatment 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,[62] and each was offered at a price of 75 cents. The purpose of this "mixed" treatment is to investigate the concern that the low probability with which any given choice problem was implemented in treatment R influenced purchase frequencies (e.g., if subjects treated the "real" choices as hypothetical).

Treatment HCT (28 subjects): Subjects performed that same task as in treatment 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.[63]

Treatment HL (28 subjects): Subjects performed the same task as in treatment 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 treatment 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).

Treatment HV (28 subjects): Subjects performed the same task as in treatment 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).

Treatment 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 treatments HWTP and L (below).[64]

Treatment 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 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.

Treatment N (28 subjects): Subjects indicated whether each potential outcome would

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

[63] 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).

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

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.

Treatment 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 treatments L and HWTP.

Treatment S (29-38 subjects):[65] 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.

D.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
- the fraction of respondents choosing purchase in the hypothetical choice question following the cheap talk script

[65] 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 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 Treatment Group S (see Appendix D.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 Treatment 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.

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

We added a "mixed" treatment, 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" treatment, 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" treatment, and estimated a logit regression relating the purchase decision to a set of 15 product dummies as well as a "mixed choice treatment" dummy. If the "real choice treatment" subjects viewed their choices as real, the coefficient for the "mixed choice treatment" dummy should be zero; if they viewed those choices as partially hypothetical, then the "mixed choice treatment" 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" treatment than in the "real choice" treatment, which is inconsistent with the hypothesis that participants in the

"real choice" treatment were more inclined to view their choices as hypothetical than were participants in the "mixed choice" treatment.

D.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.

E Microfinance Application

E.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.

E.2 Mean squared error with measurement error in covariates

In Section 6.1, 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 4 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.[66]

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.[67] 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 other settings but not setting 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 $\boldsymbol{H}_j^{r,B}$. The second term removes the effect of this error on the estimated mean squared error.

[66] 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.

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