NBER WORKING PAPER SERIES

CAUSAL INFERENCE FROM HYPOTHETICAL EVALUATIONS

B. Douglas Bernheim
Daniel Björkegren
Jeffrey Naecker
Michael Pollmann

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

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2021, Revised December 2024

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Causal Inference from Hypothetical Evaluations
B. Douglas Bernheim, Daniel Björkegren, Jeffrey Naecker, and Michael Pollmann
NBER Working Paper No. 29616
December 2021, Revised December 2024
JEL No. C13, D12

ABSTRACT

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

B. Douglas Bernheim
Department of Economics
Stanford University
Stanford, CA 94305-6072
and NBER
bernheim@stanford.edu

Daniel Björkegren
Columbia University
dan@bjorkegren.com

Jeffrey Naecker
Google, Inc.
1600 Amphitheatre Pkwy
Mountain View, CA 94043
jnaecker@google.com

Michael Pollmann
Duke University
213 Social Sciences
419 Chapel Drive
Box 90097
Durham, NC 27708-0097
michael.pollmann@duke.edu

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

1 Introduction

This paper proposes a new method for inferring the causal effects of a treatment, such as a price or policy intervention, on choices. The essence of the empirical strategy is to exploit stable relationships between choices and various types of hypothetical evaluations.

The motivation for our approach relates to the advantages and limitations of existing methods involving *stated preferences* (for reviews see Shogren 2006; Carson 2012). Imagine asking people, hypothetically, what they would choose under various conditions, and using these responses to compute treatment effects. If hypothetical choices were simply noisy measures of real choices, then this approach would offer many advantages.[1] Unfortunately, hypothetical choices are systematically biased measures of actual choices (List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005).[2] 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 approach combines the use of hypothetical responses which, though biased, can be observed for actual and counterfactual treatment states, with observational data on choices, which are observed for the actual treatment state. We consider hypothetical responses that aggregate underlying motivations (such as stated preference and hypothetical choices), as well as a variety of responses that capture more specific motivations (such as temptation or social image), including those that may influence the direction and magnitude of hypothetical bias. We estimate the predictive relationship between real choices and hypothetical responses in observational data, and then use that relationship to infer the effects of counterfactuals.

Because the method does not require treatment variation, it can recover treatment effects for novel interventions that have yet to be implemented (e.g., a policy proposal), or that appear in limited contexts (e.g., an innovative policy adopted by a small number of jurisdictions). It can also recover treatment effects when treatment assignment is random, and—under assumptions that are plausible in some applications—when treatment assignment is correlated with factors that influence the outcome. In the latter cases, it does not rely on external instruments or discontinuities, neither of which may be readily available in any given application. The method also offers other advantages. Specifically, it can (1)

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

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

recover variation in treatment effects for specified subgroups of treated or untreated units (i.e., not merely local average treatment effects or LATEs), (2) determine how treatment effects vary with complex attributes of the treatment that are not easily reducible to a small collection of variables (e.g., features of photos or text), and (3) improve statistical precision. In addition to developing the method conceptually, we offer proof of concept in two applications, one using laboratory data, the other using field data.

To be more specific, suppose we observe choices in a variety of settings, indexed by j, with a treatment state, $w \in \{0, 1\}$, assigned to each. Examples include prices set for a group of related products, or policies set for jurisdictions. The actual (aggregated) choice outcome for setting j would be $Y_j(w)$ in treatment state w. We are interested in the average treatment effect, $\tau = \mathbb{E}(Y_j(1) - Y_j(0))$. However, we observe each setting j only in the realized treatment state W_j, which may be the same for all settings. Imagine collecting hypothetical evaluations of the options available in setting j, $\boldsymbol{H}_j(w)$, for both treatment states, $w = 0, 1$ (where $\boldsymbol{H}_j(w)$ may include hypothetical choices and other evaluative responses). First, we estimate a model relating outcomes in the realized treatment states, $Y_j(W_j)$, to the corresponding hypothetical responses, $\boldsymbol{H}_j(W_j)$. Second, we use that relationship to predict average outcomes for both treatment states. The difference yields an estimate of the treatment effect. In effect, we use the estimated prediction equation to unwind the systematic biases embedded in the hypothetical responses, $\boldsymbol{H}(1)$ and $\boldsymbol{H}(0)$. We develop a simple linear estimator suitable for low-dimensional settings, as well as a machine learning estimator suitable for high-dimensional settings based on the LASSO. We also outline results for doubly robust and nonlinear estimators.[3]

As long as the predictive relationship between outcomes and hypothetical evaluations is stable, this method should yield unbiased estimates of treatment effects. We provide proof of concept by applying the method to real data involving two separate applications, one in the laboratory, the other in the field. In these applications, the method recovers measures of treatment effects that are close to ground-truth estimates, even under conditions that render standard non-structural methods inapplicable. Given this promising performance, we formalize econometric theory for the estimator, articulate conditions that would yield stability, and describe the contexts where the approach is applicable.

In our first application, we use our method to estimate the demand for various snacks as a function of prices in a laboratory setting. We ask some participants to decide whether to purchase each snack at prices \$0.25 and \$0.75. Other participants evaluate each snack

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

hypothetically along several dimensions at the low price and the high price. We simulate a data set with no price variation by restricting the choice data to a single price for all snacks. We also simulate endogenous price variation by selecting observations in a manner that introduces correlation with demand. In both of these cases, the estimates of treatment effects based on our method are close to ground truth estimates based on actual purchase decisions for each snack at both prices (which are observable in a laboratory setting).

In our second application, we use our method to assess the effects of matching provisions on lending through a microfinance platform. The observational data tell us the speed at which each borrower profile attracted funding and whether a third party offered matching funds. We gathered hypothetical data by asking Amazon Mechanical Turk workers to assess these profiles in both the matched and unmatched states. Estimates of treatment effects based on our method are once again close to ground truth, which we inferred from a controlled experiment. In contrast, a standard OLS regression of funding velocity on treatment yields a biased estimate, and no suitable instruments are available.

In addition to documenting the accuracy of our approach in two settings, our analysis of these applications illustrates how the method can shed useful light on the heterogeneity of treatment effects. It also shows that our approach can yield gains in precision even in cases where standard approaches are feasible.

To be clear, our method also has limitations. As we explain, the assumptions that justify it are potentially problematic in applications with particular features—for example, those for which it is difficult to depict decision problems comprehensively in a survey, or to survey populations that sufficiently resemble the decision makers. It also requires collecting hypothetical responses in addition to observational data on choices for a variety of decision settings. Nonetheless, in some settings the approach may provide a reliable and cost-effective alternative to field experiments, or it may complement field experiments by offering a low-cost method for exploring large varieties of treatment possibilities before committing to a particular version.

The paper is organized as follows. The next section describes our approach. Section 3 covers the laboratory application, and Section 4 covers the field application. Section 5 provides formal foundations and discusses the characteristics of appropriate (and inappropriate) applications. Section 6 extends the theory to applications with endogeneity. Section 7 concludes.

2 Method

2.1 The problem

We are interested in the effect of some treatment $w \in \{0, 1\}$ on choices made in various settings indexed by j.[4] For each setting j, the potential outcome $Y_j(w)$ represents an aggregation of people's choices (a sum or average). The objective is to estimate the average treatment effect (ATE):

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

where the expectation is taken over a population of settings.

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

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

Product demand. The analyst seeks to estimate price elasticities for a collection of products (alternatively, for the same product across different markets), but observes no variation in price. Here, settings correspond to products (alternatively, markets), the treatment is price, and outcomes are purchase decisions by customers.

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

2.2 Our approach

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

The standard method for using hypothetical choices. Imagine that in each setting, we ask people similar to the decision makers of interest what they would choose, hypothetically, under both treatment states. For example, we might ask them if they would hypothetically

[4] While we focus on environments with binary treatments, our method is more general.

purchase particular goods at particular prices, or donate to different appeals with or without matches. Using their responses, we could then construct the average hypothetical choice $Y_j^H(w)$ for setting j under treatment w.

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

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

where $\overline{Y^H(w)} = \frac{1}{J}\sum_{j=1}^{J} Y_j^H(w)$ is the sample average of the hypothetical choice under treatment state $w \in \{0, 1\}$ for all settings.

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

The main problem with this approach is that hypothetical choices are systematically biased (Cummings, Harrison, and Rutström 1995; List and Gallet 2001; Little and Berrens 2004; Murphy et al. 2005; Blumenschein et al. 2008). For example, people tend to overstate purchases, and they exaggerate their proclivities to take "virtuous" actions, such as donating to charities and purchasing healthy foods.[5]

Our basic approach. Our approach estimates how hypothetical evaluations relate to real choices, and then uses that relationship to undo the biases in hypothetical choices. We consider multiple types of hypothetical evaluations, denoted by a (row) vector $\boldsymbol{H}_j(w)$ in setting j, which may include (but is not limited to) hypothetical choices $Y_j^H(w)$. The simplest variant of our approach has two steps.

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

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

$$Y_j = \boldsymbol{H}_j\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \epsilon_j, \tag{1}$$

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

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

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

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

Because our method uses hypothetical evaluations as *predictors* rather than as *measures* of choices, we are free to use any subjective response that aids prediction. $\boldsymbol{H}_j$ can thus include not only hypothetical choices (which are aggregates of multiple underlying motivations), but also measures of specific motivations, such as the extent to which a given option satisfies a desire for health, as well as measures that may predict the direction and magnitude of hypothetical choice bias, such as whether a given option is considered socially virtuous.

Extensions involving machine learning. In applications, the number of potential predictors can be large, particularly if one seeks to include transformations such as quadratic terms (especially for hypothetical evaluations that employ arbitrary scales) and interactions (e.g., because social approval mediates the response to anticipated pleasure). Linear regression estimators may then overfit, and machine learning estimators may perform better. Appendix C.3 describes an approach similar to LASSO exploiting linearity and sparsity in high-dimensional hypothetical evaluations building on approximate residual balancing (ARB, Athey, Imbens, and Wager 2018). In Appendix C.4, we provide a doubly robust moment condition for estimation using arbitrary machine learning methods.

Stability. For the approach to work, the relationship in equation (1) must be sufficiently stable across settings and treatment states. We explore approaches to assessing stability in our two empirical examples, and formalize underlying assumptions and diagnostics in

Section 5. We also propose microfoundations for the stability assumption, which involve two premises: first, mental states fully determine choices (and in that sense are "sufficient statistics" for choices); second, mental states likewise fully determine (mis-) reporting of mental states. Under the first premise, it should be possible in principle to predict choices accurately based on mental states by recovering the stable relationship between these variables. In effect, if the *source* of variation in mental states is unimportant, then we can use mental-state variation that arises from differences between *settings* to infer the effects that would arise from different *treatment states*. While mental states are unobservable in practice, the second premise implies that they are also sufficient statistics for hypothetical evaluations. And if hypothetical evaluations span the mental states, that stable relationship is invertible. Stability of the relationship between choices and hypothetical evaluations is therefore plausible.

Overlap and linearity. Our method is most applicable when the range of variation for the hypothetical evaluations for the counterfactual treatment states, $\boldsymbol{H}_j(1 - W_j)$, sufficiently overlaps with the range of variation for the realized treatment states, $\boldsymbol{H}_j(W_j)$. These ranges tend to overlap when the effect of the treatment on evaluations is not too large relative to other sources of variation in evaluations across settings. When this spanning condition is satisfied, our method works better in practice (as demonstrated in Section 3.2.2), and does not require functional form assumptions (as shown in Appendix C.3). Caution is warranted when using functional form assumptions such as linearity (as in equation (1)) to extrapolate the relationship between y and $\boldsymbol{H}$ to values $\boldsymbol{h} = \boldsymbol{H}_j(1 - W_j)$ substantially outside the observed range of variation for $\boldsymbol{H}_j(W_j)$.

Methodological precursors. The literature on stated preferences and contingent valuation methods (SP/CVM) includes attempts to correct hypothetical bias by "fixing" the elicitation protocol (e.g., Cummings and Taylor 1999, Jacquemet et al. 2013, and Blamey, Bennett, and Morrison 1999, and Loomis, Traynor, and Brown 1999). For our method, there is no need to assume that any protocol yields an unbiased prediction. Instead, each produces a potentially useful predictor.

Other portions of the SP/CVM literature study statistical relationships between real and hypothetical choices (e.g., Blackburn, Harrison, and Rutström 1994, Fox et al. 1998, List and Shogren 1998, List and Shogren 2002, and Mansfield 1998). Instead of treating the decision problem as the unit of observation and relating choice distributions to the

problem's (subjective) characteristics as in our approach, these "calibration" studies treat the individual as the unit of observation and relate hypothetical bias to socioeconomic characteristics within a single decision setting. Because hypothetical bias is context-specific (List and Shogren 1998, List and Shogren 2002, Ajzen, Brown, and Carvajal 2004, and Johansson-Stenman and Svedsäter 2012), those individual-level relationships do not reliably transfer from one setting to another.[6] Calibration studies also potentially suffer from cross-contamination between each subject's real and hypothetical decisions. Our problem-level focus allows us to avoid cross-contamination by eliciting real and hypothetical choices from different subjects. Finally, calibration studies focus on one hypothetical question at a time. Our approach leverages the information contained in responses to multiple questions simultaneously.

Our method is also related to demand estimation approaches that augment real choices with additional data such as the alternative a consumer would hypothetically choose if the real choice were unavailable (Berry, Levinsohn, and Pakes 2004) or measures of relatedness gathered from surveys (Magnolfi, McClure, and Sorensen 2022). There is also related work in marketing (Infosino 1986; Jamieson and Bass 1989; Morwitz, Steckel, and Gupta 2007), political science (Jackman 1999; Katz and Katz 2010), and neuroeconomics (Smith et al. 2014). See Appendix B for more discussion of these connections.

Other related methods. Appendix C.9 describes connections to linear factor models, synthetic controls, and statistical surrogates.

2.3 Potential advantages

When our method is applicable, it offers several potential advantages.

Effects of treatments that have not been implemented. Because our method does not require *any* variation in assigned treatment states, it enables the estimation of causal effects for novel treatments that have yet to be implemented. Intuitively, if the hypothetical evaluations in the untreated (baseline) state mostly span the range of variation under

[6] Blackburn, Harrison, and Rutström (1994) provide somewhat mixed evidence on portability, but their analysis is limited to two goods. Unlike calibration studies, meta-analyses (e.g., List and Gallet 2001, Little and Berrens 2004, and Murphy et al. 2005) attempt to account for the variation in hypothetical bias across contexts and goods, but mainly as a function of coarse features of the goods (public versus private) and experimental methods.

treatment, we can infer what people would choose on average in a treated setting by examining choices in untreated settings that evoke similar hypothetical evaluations.

Effects of endogenous treatments. The approach also provides novel routes to identification when observed treatment is assigned endogenously. Because Step 2 uses data on *both* treatment states for every setting, treatment endogeneity can only bias the estimate by distorting the value of $\boldsymbol{\beta}$ obtained in Step 1. We derive $\boldsymbol{\beta}$ from the relationship between the observed outcome Y_j and hypotheticals in the corresponding treatment state, $\boldsymbol{H}_j = \boldsymbol{H}_j(W_j)$. If treated settings would have had higher (or lower) outcomes regardless of treatment, the method will attempt to explain that difference in outcomes not through differences in treatment states, as with a quasiexperimental estimator, but through differences in the hypotheticals associated with the corresponding treatment states. But the treatment is only one source of variation for the hypotheticals. There is also natural variation across settings: some settings simply have higher evaluations than others, regardless of treatment. Because this "ambient" variation helps identify the relationship between outcomes and hypotheticals, it can dilute any endogeneity problem. Endogeneity is only a significant problem when too much of the variation in hypotheticals comes from the treatment—that is, when the settings are not sufficiently diverse apart from the treatment. Conversely, when there is sufficient ambient variation in the hypotheticals, the bias from treatment endogeneity becomes small. Our method also works well in settings where hypothetical evaluations plausibly span most of the outcome-relevant portion of the information used to select the treatment. We formalize these points in Section 6, where we also introduce methods of bounding and correcting remaining biases.

Heterogeneous treatment effects. Treatment effects commonly vary across units (here, across settings). Standard observational methods identify treatment effects only for the specific units that are affected by an instrument or discontinuity (the Local Average Treatment Effect (LATE) among compliers; Imbens and Angrist 1994). Using our methods, one can estimate the average treatment effect (ATE) for any subgroup of settings $\mathcal{S}$ (defined according to values of our conditioning variables $\boldsymbol{H}_j(1), \boldsymbol{H}_j(0)$, and $\boldsymbol{X}_j$) by calculating $\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\left[\hat{Y}_j(1) - \hat{Y}_j(0)\right]$. This calculation is possible because both treated and untreated hypothetical responses are observed for all settings.

Encoding and evaluating nuanced features of treatments. Standard causal inference methods are typically constrained to use coarse definitions of treatments and settings. For

example, the literature on organ donation focuses on a single feature: whether people are invited to opt in or opt out (Kessler and Roth 2012). Our approach allows one to estimate how treatment effects vary with more nuanced features, such as the wording and placement of an organ donation question, or the photo and text used alongside a microfinance profile. The analyst need not encode those features manually; instead they can simply assess hypothetical responses to treatments that differ only with respect to a given feature, and apply the estimated prediction equation. Our methods can therefore complement field experiments by allowing analysts to explore the treatment design space at far lower cost by gathering hypothetical responses, and then focusing on the most promising designs.

Precision. Compared with standard methods, our approach can improve the precision of estimated treatment effects even when randomized treatment variation is available, particularly when treatment groups have unbalanced sizes. Because we estimate a single model of the outcome as a function of hypothetical evaluations using all settings (treated and untreated), and then use hypothetical data in both treatment states to predict outcomes for every setting, imbalance has no direct impact on the precision of our method.

3 Application: Snack Demand

We first demonstrate our approach by estimating price sensitivities for a family of goods in a laboratory setting. Study participants make simple purchase decisions for a large collection of familiar snack foods. The treatment states $w \in \{0, 1\}$ correspond to prices of \$0.25 or \$0.75, respectively. $Y_j(w)$ denotes aggregate demand for good j at the price corresponding to w. The treatment effect of interest is either the average price response $\frac{1}{J}\sum_{j=1}^{J}[Y_j(1) - Y_j(0)]$, or the responses for individual goods.

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

3.1 Procedures and data

Each of 365 subjects was assigned to one of several groups, described below.[7] Subjects were told that their sessions consisted of two stages. The first involved a computer-based choice or rating task lasting roughly 30 minutes. The second was a 30-minute waiting period. Subjects were asked not to eat anything during the waiting period unless a snack was provided (according to the rules).

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

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

3.1.1 Real choices

The subjects who made real choices were informed that we would select one decision at random and implement it during the 30-minute waiting period.[8] Although the chance of implementing any given choice was low, differences between real and hypothetical purchase frequencies were substantial, and in the expected direction.[9]

In observational data we might observe such demand at a single price for each good,

[7] We conducted the experiment at the Stanford Economic Research Laboratory (SERL) between November 15, 2010, and October 2, 2012. Stanford University's IRB reviewed and approved the protocol. The participation fee ranged from \$20 to \$30. We adjusted the fee upward when the response rate to our subject solicitation was low, and downward when it was high. Sessions took place in mid-afternoon, when subjects are typically hungry.

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

[9] Real purchase frequencies were not significantly different in a group of participants whose odds of implementation were one in 5 decisions rather than one in 378. See Appendix D.3 for details. It is not surprising that participants in the "real choice" group viewed their choices as real: they had as much at stake as someone making a single purchase decision (because they knew we would definitely implement one choice), and taking the task less seriously did not reduce the subject's time commitment. Notably, similar conclusions were reached by Carson, Groves, and List (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.

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

3.1.2 Hypothetical evaluations

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

Several groups made hypothetical choices. The literature on stated preferences explores a variety of protocols for eliciting such choices. We employed multiple protocols, each with a separate group. The "standard" protocol mimicked the real choice protocol, except that no choice was implemented. A second protocol employed a "cheap talk" script (as in Cummings and Taylor 1999) that encouraged subjects to take the hypothetical choices seriously,[10] 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 2011a), and a fifth elicited hypothetical willingness-to-pay (WTP) rather than Yes/No responses.

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

3.1.3 Patterns of real and hypothetical choices

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

Hypothetical choices overstate demand: when asked hypothetically, demand is nearly 7 percentage points higher (equivalently, 28% higher: 31% vs. 24%) across all item-price pairs, and we reject the absence of bias ($p \leq 0.001$). Moreover, hypothetical demand

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

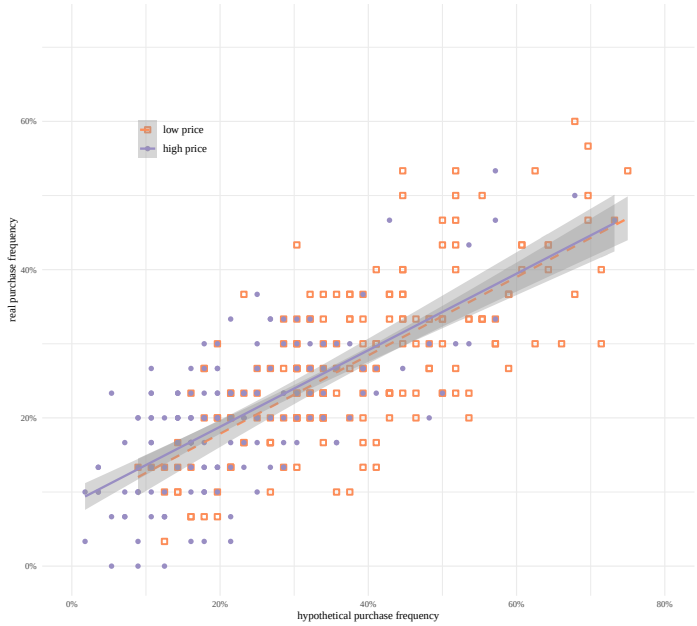


Figure 1: Real vs. Hypothetical Choices

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 Appendix Figure A2, we show that the curves are approximately linear and similarly overlap when using nonparametric regression.

exceeds the real demand for 70% of item-price pairs. Additionally, hypothetical demand is more variable, with more than twice the variance of real demand across all item-price pairs; see Appendix D.4 for additional analysis of this difference. A possible explanation is that, when answering hypothetical questions, people naturally exaggerate the sensitivity of their choices to characteristics and conditions.

Although hypothetical demand is a poor predict*ion* of real demand, it is strongly correlated with real demand, and consequently may be a useful predict*or*. Figure 1 illustrates this relationship, with the demand for each item shown as orange squares when priced at $0.25, and as purple dots when priced at $0.75. The relationship between hypothetical and real demand is systematic, and, helpfully for our purposes, stable between treatments.[11]

3.2 Effect of an unseen counterfactual

Our method can reveal treatment effects in applications for which there is no real-world variation in the treatment of interest. The reliability of the estimate will depend on how well

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

the relationship between choices and hypothetical responses (equation 1) extrapolates into the unseen treatment state. We articulate theoretical conditions under which extrapolation is accurate in Section 5.

Estimators based solely on observed choices are infeasible for this task. Panels (a) and (b) of Table 1 show estimates of treatment effects, along with standard errors, for approaches that use hypothetical evaluations. In both panels, Column (1) reports the ground truth estimate, that increasing the price from \$0.25 to \$0.75 changes the proportion of subjects buying the average snack by -0.075. We first compare this ground truth to estimates that treat various measures of hypothetical choices discussed in the literature as predictions; then we do the same for estimates based on the method proposed in this paper.

3.2.1 Estimators that treat hypothetical choices as predictions

Columns (2) through (6) of Table 1, panel (a), report the difference between hypothetical choice frequencies, elicited through a specified protocol, at the two prices. These estimators do not require data on real choices.

Treating standard hypothetical choices as predictions (i.e., estimating the effect as the mean difference in hypothetical choices, Column (2)) yields an estimated effect of -0.159, more than twice ground truth. This discrepancy reflects significant hypothetical bias.

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

As shown in Columns (3)–(6), 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 -0.091, which is closer to the true effect. However, had we not known the ground truth, we would have had no basis for selecting the dichotomization threshold used for this estimate over other thresholds, which yield less accurate estimates. Moreover, it appears that the improvement is accidental, and does not reflect more informative responses. In particular, the final two rows of the first panel report correlations between

Table 1: Estimating Treatment Effects without Variation in Treatment

(a) Hypothetical as Prediction

	Ground Truth	Hypothetical as Prediction				
		Diff. in Hypotheticals				
	(1)	(2)	(3)	(4)	(5)	(6)
Estimated effect of high price	-0.075	-0.159	-0.188	-0.129	-0.091	-0.266
	(0.004)	(0.006)	(0.007)	(0.006)	(0.005)	(0.009)
Hypotheticals:						
. . . hypothetical choice		X				
. . . cheap talk			X			
. . . intensity as choice				X		
. . . vicarious as choice					X	
. . . WTP as choice						X
Sample size (outcome)	189 ($\times 2$)	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

(b) Hypothetical as Predictors

	Ground Truth	Hypothetical as Predictors				
		Low Dimensional				
	(1)	(2)	(3)	(4)	(5)	(6)
Observing all snacks at high price						
Estimated effect of high price	-0.075	-0.082	-0.086	-0.076	-0.050	-0.078
	(0.004)	(0.008)	(0.012)	(0.012)	(0.007)	(0.013)
Observing all snacks at low price						
Estimated effect of high price	-0.075	-0.084	-0.101	-0.069	-0.048	-0.147
	(0.004)	(0.008)	(0.009)	(0.008)	(0.006)	(0.016)
Hypotheticals:						
. . . hypothetical choice		X				
. . . cheap talk			X			
. . . intensity as choice				X		
. . . vicarious as choice					X	
. . . WTP as choice						X
Sample size (outcome)	189 ($\times 2$)	189	189	189	189	189

Estimates of the effect of the high price (vs. low price) on the real purchase frequency. Analytical standard errors are in parentheses.

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

3.2.2 Our approach

In effect, our method predicts outcomes for a given setting in an unseen treatment state by extrapolating from observed outcomes in settings that induce similar motivations (as measured by hypothetical responses) in the prevailing treatment state. We check whether the distribution of evaluations over settings in the prevailing treatment state overlaps with the corresponding distribution in the unseen treatment state in Appendix D.5. For the various hypothetical choice variables (choice, cheap talk, own choice likelihood, vicarious choice likelihood), overlap between the distributions at the two prices is reasonably complete. It is less complete for the dichotomized WTP choice variable, so we exclude that variable from our multivariate specifications throughout. Further analysis of the WTP choice variable highlights the dangers of extrapolating beyond the range of observed variation.

Treatment effect estimates. In panel (b) of Table 1, we exhibit estimators based on univariate models that relate the outcome to each hypothetical variable individually. The table distinguishes between two cases, depending on whether we allow the estimator to observe real choices at the high price or at the low price. The estimates in Columns (2)–(5) range from -0.048 to -0.101. The dichotomized WTP choice yields an accurate estimate when all snacks are observed at the high price, but shows a larger bias when all snacks are observed at the low price (Column (6)), which is anticipated based on the violation of overlap we documented. Overall, using even a single hypothetical choice variable as a predictor shows promise for estimating the effect of price changes in the absence of observed variation in prices.

Table 2: Estimating Treatment Effects without Variation in Treatment (multivariate)

	Ground Truth	Our Method: Hypotheticals as Predictors				
		Low Dimensional		High Dimensional		
	(1)	(2)	(3)	(4)	(5)	(6)
Observing all snacks at high price						
Estimated effect of high price	-0.075	-0.082	-0.075	-0.073	-0.084	-0.077
	(0.004)	(0.010)	(0.010)	(0.012)	(0.005)	(0.012)
	[0.004]	[0.009]	[0.010]	[0.017]	[0.016]	[0.024]
Observed at high price	All			All		
Observed at low price	All			None		
Observing all snacks at low price						
Estimated effect of high price	-0.075	-0.100	-0.096	-0.098	-0.094	-0.096
	(0.004)	(0.008)	(0.008)	(0.009)	(0.004)	(0.013)
	[0.004]	[0.007]	[0.007]	[0.015]	[0.014]	[0.017]
Observed at high price	All			None		
Observed at low price	All			All		
Controls			X	X	X	X
Hypotheticals:						
. . . all hypothetical choices (excl. WTP)		X	X	X	X	X
. . . detailed hypothetical eval. (excl. WTP)					X	X
2nd order + interactions				X		X
Sample size (outcome)	189 ($\times 2$)	189	189	189	189	189

Estimates of the effect of the high price (vs. low price) on the real purchase frequency. Analytical standard errors are in parentheses; bootstrap standard errors in square brackets are based on 1,001 bootstrap samples. Analytical standard errors for the high-dimensional specifications treat average evaluations as fixed, whereas all other standard errors consider estimation error in the average evaluations across snacks. For results including WTP see Appendix Table A1.

Our method may perform even better when it employs multiple hypothetical covariates that more comprehensively span motivations. Table 2 explores this possibility. Column (1) reproduces the true average treatment effect. For our method, we show results based on several specifications of the prediction model. For Column (2), we use the four hypothetical choice variables together (omitting the dichotomized WTP variable).[12] For Column (3), we add eight conventional controls that are available in standard datasets (physical characteristics, including grams per serving and seven measures of nutrients). 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 Appendix C.3. The first of these (Column (4)) includes the various hypothetical choice variables and eight physical characteristics, as well as second order and interaction terms. The second specification (Column (5)) 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 (6)) adds a complete set of second-order and interaction terms.

The estimates are close to the true average effect when all snacks are observed at the high price (between -0.073 and -0.084). Even with the dichotomized WTP choice variable omitted, accuracy is somewhat lower when all snacks are observed at the low price; possibly the failure of overlap for that variable reflects considerations that somewhat compromise stability more generally. Nevertheless, as we show in the next section, our method does a good job of capturing the heterogeneity of treatment effects in *both* cases.

3.3 Heterogeneity in treatment effects

Using our method, hypothetical evaluations 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.

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

[12] With WTP included, when all snacks are observed at the high price, estimates are similar; when all snacks are observed at the low price, including WTP causes most estimates to be further from ground truth; see Appendix Table A1. The reason is that, when snacks are observed at the high price, we do not need to extrapolate as much to predict demand at the low price, as shown in Figure A3(a).

- *R^2 for a regression of τ_j on $\hat{\tau}_j$:* This statistic measures the fraction of the variation in true treatment effects that the estimates capture.

- *Mean squared error (MSE) relative to predicting with the ATE ($\frac{\text{mean}((\tau_j-\hat{\tau}_j)^2)}{\text{mean}((\tau_j-\tau)^2)}$):* This statistic encompasses overall accuracy and precision.

- *Calibration coefficient:* This measure is the slope coefficient in a regression of τ_j on $\hat{\tau}_j$. The ideal coefficient is unity: in that case, the expectation of the actual treatment effect increases unit for unit with the predicted treatment effect.

- *Simulated profit:* We simulate a producer who estimates heterogeneous price sensitivity for each snack j in order to set prices w_j^*. We report the gain in average profit, relative to setting prices at random, as a fraction of the maximum possible gain achieved by optimal pricing, $\frac{\bar{\pi}(\boldsymbol{w}^*)-\bar{\pi}(\boldsymbol{w}^{\text{random}})}{\bar{\pi}(\boldsymbol{w}^{\text{optimal}})-\bar{\pi}(\boldsymbol{w}^{\text{random}})}$, where $\bar{\pi}$ is the average profit as a function of J prices. For details see Appendix D.6.

Results. Results appear in Figure 2. In this section, we focus on environments where there is no variation in treatment and ones where treatment is assigned randomly, 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.

With conventional approaches based solely on observational data, no treatment effects can be estimated when there is no variation in assigned treatment.

When treatment is assigned randomly, one can use the difference-in-means estimator as a benchmark (row 1).[13] Because this estimate does not vary with j, R^2 and the calibration parameter are both zero. Even so, if the available covariates have little explanatory power, this simple estimator may perform well relative to alternatives in terms of MSE and simulated profits by virtue of its parsimony.

With random treatment variation, 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. This approach would be infeasible in applications because it would require observations of each unit in both treatment states. It is nevertheless of interest because it benchmarks the greatest amount of heterogeneity one might hope to capture by conditioning on the physical characteristics

[13]The difference-in-means estimator is $\hat{\tau}_j \equiv \hat{\tau} = \frac{1}{\sum_{j'=1}^{J} W_{j'}} \sum_{j'=1}^{J} W_{j'} Y_{j'} - \frac{1}{\sum_{j'=1}^{J} 1-W_{j'}} \sum_{j'=1}^{J} (1-W_{j'}) Y_{j'}$.

linearly.[14] In rows 3, 4, and 5, we also consider three conventional estimators that are feasible in 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.

Alternately, a standard hypothetical approach would use data solely on hypothetical evaluations to estimate heterogeneous treatment effects (row 6). We estimate the treatment effect for each item as the difference between hypothetical choice frequencies at the two prices, elicited with the standard protocol. This estimate is feasible even when treatment has no real world variation. Notably, this method yields substantially higher R^2 than conventional methods, highlighting that hypothetical evaluations contain much useful information. However, hypothetical choice bias leads to very large mean squared error.[15]

Our method (rows 7-10) combines hypothetical evaluations and observations on choices, which allows it to also describe much heterogeneity while removing the hypothetical choice bias, thereby reducing mean squared error and further improving calibration and profit. These rows of Figure 2 show results for the variant that employs hypothetical choices and physical characteristics as predictors (i.e., the same variant as in Table 2 Column (3)). The method attains similar performance regardless of whether treatment assignment has no variation or is randomized. For row 7, effects are estimated based on a sample where all snacks are observed at the high price. For row 8, all snacks are at the low price. Row 9 uses randomized assignment. Row 10 demonstrates that our method achieves similar performance when we simulate endogenous treatment assignment; we discuss it in Section 3.5.

Our method performs substantially better across the board than the three feasible conventional estimators that do not use hypothetical evaluations. It also easily surpasses the infeasible benchmark with respect to all metrics other than calibration, for which that benchmark mechanically achieves a coefficient of 1 regardless of which covariates are included. The comparisons to the infeasible estimator imply that hypothetical evaluations contain substantially more information about variation in treatment effects than physical characteristics in our setting, even though unadjusted hypothetical choices do not reflect that variation accurately. Because our method does not require an intervention, it can

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

[15] The other hypothetical variables mostly perform worse except with respect to mean squared error, as shown in Appendix Figure A4.

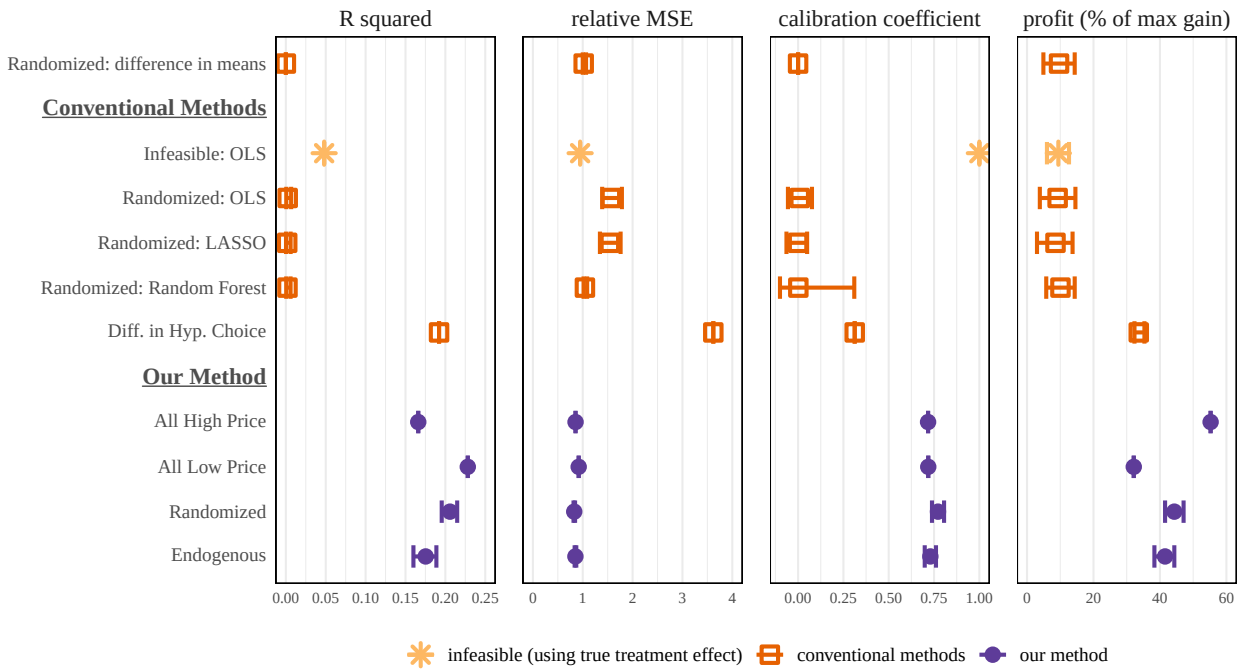


Figure 2: Treatment Effect Heterogeneity

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. For the calibration coefficient, the lower boundary of the first quartile for the random forest estimator is -0.32, but is shown in the figure as -0.1 because the axis is truncated. See Appendix Table **A2** for additional specifications.

enable analysts to recover heterogeneous treatment effects even when they lack the power to intervene before a broader roll-out. This feature may be particularly valuable in settings where one wishes to target the treatment at those who are likely to benefit most.

3.4 Gains in precision

Our method may also yield more precise estimates even when treatment is randomly assigned. Most notably, the performance of standard methods deteriorates when the fraction treated is far from half, while our method maintains good performance even if few of the observations are treated (or none, as in Section 3.2). Also, in applications, it may be easier to improve precision by collecting hypothetical responses, rather than by expanding an experimental sample.

We explore these issues in an environment with random treatment assignment. Fixing the fraction of snacks observed at the higher price, we simulate uncertainty in treatment assignment by randomly dividing the snacks into high-price and low-price subsets of fixed

sizes. We generate 1,001 such random samples and report performance metrics for different estimators as functions of the fraction of snacks observed at the high price in Figure 3.[16] We consider two standard approaches, difference-in-means and ARB (with conventional controls as well as second-order terms and interactions). We also use two variants of our method, the univariate specification using the standard hypothetical choice and the high dimensional specification from Column (8) of Table 2.

The standard deviations of our estimators are substantially smaller than those of the conventional estimators, especially when the proportion treated is far from half, as shown in the first panel of Figure 3. The standard deviation of the difference-in-means is U-shaped in the fraction of treated observations. When half of the sample is treated, the (median) standard error of the difference-in-means is more than twice that of the univariate hypothetical choice estimator.[17] To achieve the same standard error for the difference-in-means as for our univariate hypothetical choice specification with 189 snacks, one would need a randomized experiment with over 800 snacks. As the sizes of treated and untreated subsamples become less balanced, conventional estimators dramatically lose precision because the smaller of the treatment and control groups dominates the variance. In contrast, the precision of our low-dimensional estimator is largely independent of the proportion treated. The reason is that the first step of our method pools all observations, and the second uses the hypothetical evaluations for *both* treatment states for every snack.

In this application, a smaller standard deviation comes at the cost of a small bias (Figure 3 second panel), but our estimators attain lower root-mean-squared error, irrespective of the treatment's prevalence (Figure 3 final panel). The difference-in-means is unbiased by design, and hence its root-mean-squared error is equal to its standard deviation. The univariate hypothetical choice method entails a slightly larger bias, but the reduction in variance more than compensates in terms of root-mean-squared error.

3.5 Estimation under endogenous treatment assignment

Our method is also potentially suitable for settings in which treatment selection is correlated with measured or unmeasured factors that influence the outcome. Appendix D.7 mimics an observational dataset for markets in which firms assess the demand for each snack item through surveys and use that information to set price. Regressions that control for

[16] 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.

[17] For standard errors and probability of coverage of confidence intervals, see Appendix Figure A5.

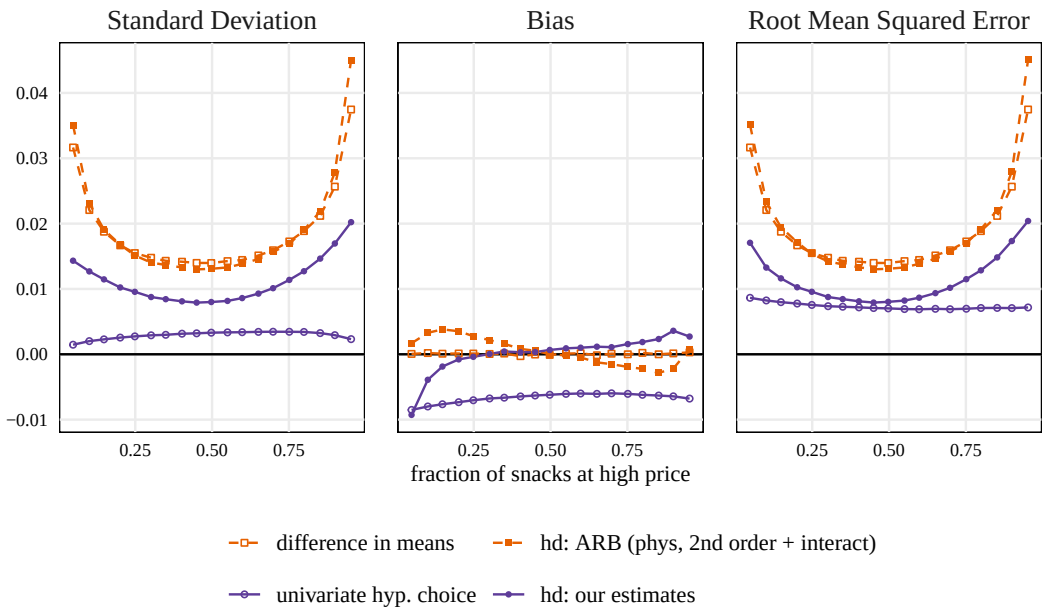


Figure 3: Performance of Estimators by Fraction Treated

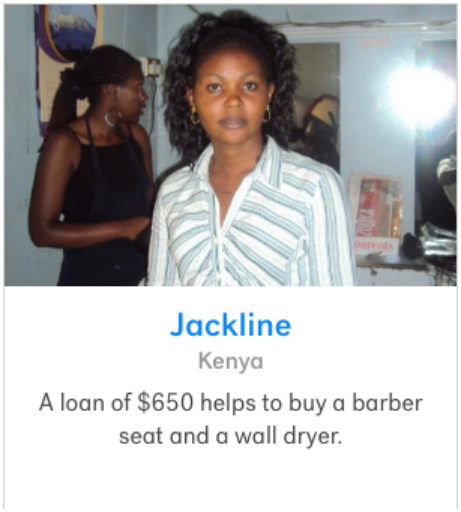
The horizontal axis measures the fraction of snacks randomly assigned to the high-price treatment state. At the boundaries of the interval, only our estimators are well-defined (see also Section 3.2), and the standard deviation across assignment realizations is mechanically zero because there is only one possible assignment. For additional specifications see Appendix Figures A6 and A7.

the treatment and standard covariates yield substantial endogeneity bias (yielding an estimated treatment effect of -0.028); however, our method continues to yield estimates close to ground truth (-0.070 to -0.081, relative to ground truth -0.075). The next section demonstrates that it also works in a field application with endogenous assignment, and Section 6 explores formal reasons for favorable performance under endogeneity.

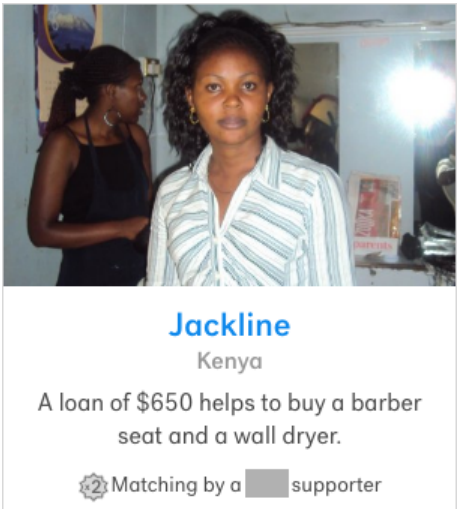
4 Application: Microfinance Contributions

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

We focus on a large microfinance crowdsourcing website, which displays profiles of potential borrowers and allows website visitors to contribute to their loans. Contributors



(a) Unmatched profile



(b) Matched profile

Figure 4: Loan Profiles with Matching Indicator

are typically socially minded individuals in developed countries; borrowers are typically developing country residents who request funds for various projects (such as business, agricultural, home, or health expenses). The website allows sponsors to offer matching funds based on criteria the sponsors specify. When a loan is eligible for a match, the profile prominently displays an indicator (as shown on Figure 4). For every dollar the visitor contributes, the sponsor also contributes a dollar.

In this setting, the unit of observation j is a loan profile, and the potential outcome $Y_j(w)$ is fundraising velocity for the first 24 hours after the loan appears on the website. We transformed velocity using the inverse hyperbolic sine to reduce the impact of outliers.[18] The treatment $W_j \in \{0, 1\}$ specifies whether the loan is matched by sponsors, who may be individuals or collectives such as churches or community groups. The treatment effect of interest is the impact of matching on fundraising velocity. Because sponsors can specify criteria for loan selection (for example, based on the borrower's gender, region, sector, loan size, risk, and/or number of days until expiration), matching status may be endogenous. Endogeneity arises from correlations between the preferences of sponsors and contributors.

We compare estimates from observational data using standard methods and our method, against a "ground truth" estimate based on a field experiment in which we introduce randomly assigned matches building on previous experiments in the literature (Karlan and

[18]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.

List 2007; Huck and Rasul 2011).[19]

4.1 Data

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

4.1.1 Observational data

Through a collaboration with the website, we observe 11,668 loan profiles for borrowers seeking $1,000 or less posted between October 14, 2019, and November 3, 2019 (we omit a random subsample that served as the treatment group for our experiment, as described below). We retain 9,623 profiles (82%) that were either unambiguously matched (matched for at least 90% of the first 24 hours after their initial posting) or unmatched (matched for no more than 3% of the first 24 hours). After dropping the remaining 18% of profiles, which were matched for intermediate fractions of the first 24 hours, we create a binary treatment indicator.[20] According to this indicator, 623 (6.5%) of the retained profiles were matched, so the treatment is rare. For all profiles, our data include descriptive characteristics, whether it was matched, and how quickly it raised funds.

4.1.2 Hypothetical responses

Separately, we collected responses to hypothetical questions concerning a subset of the loan profiles from 833 participants recruited through Amazon Mechanical Turk. We selected 200 unmatched and 100 matched loan profiles at random from the observational sample for our exercise, oversampling matched loans. For each participant, we selected 30 of these profiles at random, 20 drawn from the sample that was not matched on the website, and 10 from the sample that was matched. Participants initially viewed an overview page with a large collection of "thumbnail" profiles that reflected the overall prevalence of matches among active loans on the website. Then participants saw each of their selected 30 profiles either in the same treatment state as on the website or edited to appear in the opposite state (with the matching funds indicator either added or removed). We displayed 15 of the 30 selected profiles as unmatched (10 of which were actually unmatched on the website) and 15 as matched (5 of which were actually matched on the website).

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

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

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

4.1.3 Ground truth experiment

We established ground truth through an experiment. Starting on October 27, 2019, we assigned all new loan listings for borrowers seeking $1,000 or less either to a treatment group (roughly 10%) or a control group (roughly 90%).[21] We established a sponsorship account for loans in the treatment group and used it to ensure that contributions to them were matched for the first 24 hours after they appeared on the website. We stopped adding loans to our sample once the funds in the sponsorship account were depleted. The resulting treatment group includes 109 loans, and the resulting control group includes 982 loans.

Other sponsors continued to match loans during the course of our experiment. For the treatment group, the website used matching funds from our sponsorship account only if the loan did not meet the criteria set for any other active sponsorship account. Loans that would be matched irrespective of our intervention are always-takers; they correspond to matched loans in the observational sample. Loans that would not have been matched in the absence of our intervention are compliers.[22]

4.2 Treatment effects

Table 3 contains estimated treatment effects for matching provisions derived through a variety of methods. Because we ran an experiment, we can estimate a ground truth

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

[22] 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.

Table 3: Estimated Treatment Effects from Microfinance Application

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

Estimates of the effect of matching on the inverse hyperbolic sine of fundraising velocity, within the first day. Controls include dummies for gender, region, and sector. 'Avg. hypothetical eval.' includes the mean responses concerning projected quintile for fundraising velocity, contribution likelihoods (respondent and typical user), and funding allocation. 'Freq. hypothetical eval.' includes the frequency of "at least" each potential response to each hypothetical question (for instance, the frequency of respondents projecting the second or higher quintile, the third or higher quintile, etc.). '2nd order' includes quadratic terms for the mean responses and frequencies of each hypothetical response (if used). 'interactions' includes all two-way interactions between mean responses, frequencies of each hypothetical response (if used), and the controls. Analytical standard errors in parenthesis, bootstrap standard errors in square brackets.

measure of the treatment effect (Column (1)). In the experiment, the intention to match is random, so we use it as an instrument for the endogenous matching indicator. The standard instrumental variables (IV) estimate of the LATE is 1.24 (s.e. 0.33), which measures the effect on compliers (unmatched loans).

Next we attempt to recover treatment effects using only the observational data. It is difficult to unravel the structure of the process that renders matching provisions endogenous: conditioning the analysis on sponsors' potential matching criteria is impractical and potentially ineffective,[23] and good instruments are not readily available. Because the types of loan profiles that draw matching funds also tend to attract contributions, estimators that do not address this endogeneity exhibit substantial bias. The simple difference in means implies an estimated treatment effect of 2.55 (Column (2)), more than twice the

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

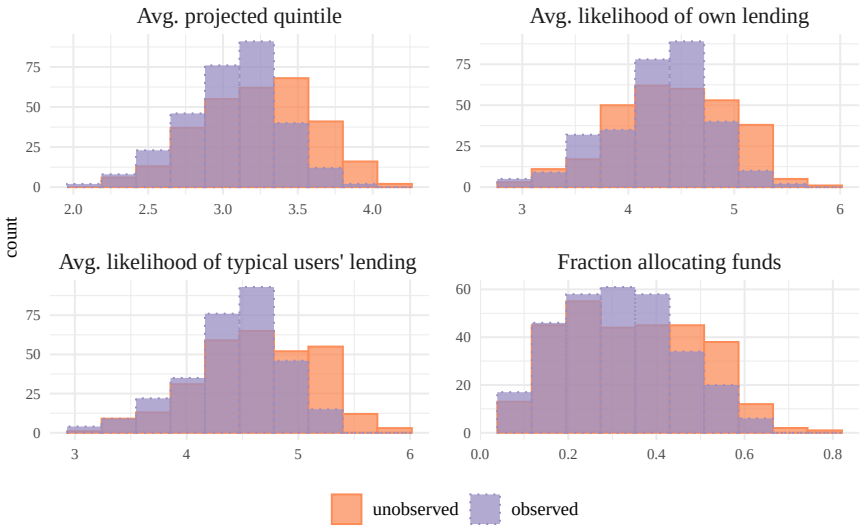


Figure 5: Overlap in hypothetical evaluations for loan $\times$ treatment states that are observed (blue) vs. unobserved (red) in the data.

ground truth. Controlling for standard covariates such as the gender and industry of the loan profile does not reduce this bias, regardless of whether we insert each factor linearly (Column (3)) or flexibly control for linear and interaction terms using ARB (Column (4)). We reject equality between each of these estimates and the ground truth.

Next we turn to estimates based on hypothetical evaluations. We first check overlap. Figure 5 shows that, for most of the evaluations of profiles in counterfactual states (red), there are indeed loans with similar evaluations in their assigned states (blue). Consequently, our method requires only modest extrapolation (for profiles and treatments that are highly desirable). We make predictions in our second step for only the unmatched loans so that our method estimates the same object as the experiment (the LATE is the average treatment effect on the control, ATC).

Our method yields estimates ranging from 0.62 to 1.80. Table 3 exhibits a low-dimensional specification that includes the average of each hypothetical evaluation linearly (Column (5)), as well as ones that add standard controls (Column (6)), squared hypothetical evaluations (Column (8)), and both (Column (9)).[24] It also includes high-dimensional specifications based on our ARB estimator that add interaction terms (Column (9)), distribution detail for each possible hypothetical response (Column (10)), and both (Column (11)). Statistical tests fail to reject the hypothesis that each of the estimates using our method coincides with the ground truth. But, crucially, our method requires only a *hypothetical* experiment.

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

4.3 Heterogeneity: treatment effects by complier group

Even in settings where one has a randomized or natural experiment with imperfect compliance, the instrumental variables estimator will recover the treatment's effect only on compliers (a LATE). In many applications, the analyst may be interested in treatment effects for other groups. For example, if we were deciding whether to eliminate the 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, if deciding whether to make a matching policy universal, we would like to evaluate its overall effect (ATE).

Our method can in principle estimate average treatment effects for any specified subgroup. We illustrate this feature in Table 4. The first row reproduces selected estimates of the LATE (also the ATC) from Table 3, including the IV estimate, as well as two measures obtained through our method (corresponding to the low and high dimensional specifications in, respectively, columns (5) and (11) of Table 3). 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 appear to differ among compliance groups. The second row shows that our estimates of the average treatment effect on the treated (ATT) is less than half the LATE/ATC for both specifications. That our method reproduces estimates close to ground truth for the LATE/ATC increases confidence that the estimate of the ATT is also reliable. Loans that are matched in practice apparently do not benefit as much from the match, presumably because they are sufficiently attractive in other dimensions to raise funds irrespective of matching. In this case, the estimated ATEs are close to the LATE/ATCs because the population of always-takers is relatively small (6.5% of the total). Nevertheless, our finding has an immediate policy implication: the microfinance platform may be able to raise more funds by inducing sponsors to match contributions to the loans they currently are not matching.

Table 4: Heterogeneity by Compliance Group in the Microfinance Application

	Experiment	Our Method		Proportion
	IV (1)	Low Dimensional (5)	High Dimensional (11)	of Observational Sample
Estimated effect of matching				
..... on compliers (LATE/ATC)	1.24 (se 0.32)	0.62 (se 0.33)	1.38 (se 0.47)	93.5%
..... on always-takers (ATT)	cannot be estimated	0.11 (se 0.19)	0.75 (se 0.38)	6.5%
... average (ATE)	cannot be estimated	0.59 (se 0.32)	1.34 (se 0.44)	100%
Test: equal effects (p-value)	—	0.03	0.27	

The first row of estimates reproduces results from Table 3, columns (1), (5), and (11) (as indicated in the column headings). Standard errors in parenthesis are based on the bootstrap.

5 Formal Results

Given the promising performance of our approach, we provide formal statistical foundations, clarify underlying assumptions, and describe the characteristics of suitable applications.

5.1 Statistical assumptions and properties

This section lists statistical assumptions that are sufficient to ensure our simple linear estimator for the ATE is consistent and asymptotically normal.

Assumption 1. Invariant mapping. *The mapping between potential outcomes and hypothetical evaluations is the same in either treatment state:*

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

Assumption 2. Linearity. *The conditional expectations of potential outcomes are linear in the predictors: for* $w \in \{0, 1\}$,

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

The simplest variant of our method requires the two preceding assumptions, which we have stated in order of increasing restrictiveness. Relaxing them is feasible but leads to more complex variants; see Appendix C.3. A third assumption governs how treatment is assigned. Here, there are several options. The properties of our method are easiest to understand when there is no variation in treatment or it is randomly assigned. Formally:

Assumption (3A)**. No treatment variation.** *In the data on real choices, there is no variation in treatment:* $W_j = 0$ *for all settings* j*, or* $W_j = 1$ *for all settings* j*.*

Assumption (3B)**. Treatment is randomly assigned.** *In the data on real choices, treatment is randomly assigned: For all settings* j*,* $\Pr(W_j = 1) = p_j = p$*.*

In Section 6, we introduce alternative assumptions that extend the method to cases where part of the variation in treatment assignment is endogenous.

The stated assumptions are sufficient to ensure that our linear estimator has the following asymptotic distribution.

Proposition 1. *Suppose the data* $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j)_{j=1}^J$ *are a random sample of independent observations and standard regularity conditions hold. Under Assumptions 1, and 2, if either Assumption 3A or 3B (or 3D introduced later) holds, the parametric estimator* $\hat{\tau}$ *is consistent for the average treatment effect* τ *and asymptotically normal:*[25]

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

5.2 Characteristics of suitable applications

In this section, we explore the plausibility of Assumption 1 and identify the types of applications that might satisfy it.

Any decision problem involves choosing from a menu of options. When someone makes a choice, their brain maps each option to a bundle of "motivational attributes" (e.g., the degree to which the option addresses hunger, social approval, and so forth). We can therefore think of the individual as choosing from a "psychological menu" containing bundles of motivational attributes (Shenhav 2024). The central premise of our approach is that if two decision problems map to the same psychological menu, the options a person would select in each problem map to the same item on that menu (or to one that is equally

[25] The formula for the variance matrix is:

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

where, for notational convenience, we denote the full sets of regressors by $\boldsymbol{Z}_j(w) = [\boldsymbol{H}_j(w),\ \boldsymbol{X}_j]$, $\boldsymbol{Z}_j = \boldsymbol{Z}_j(W_j)$, and the joint vector of their coefficients by $\boldsymbol{\delta} = [\boldsymbol{\beta}^T,\ \boldsymbol{\gamma}^T]^T$. $\boldsymbol{V}^{\text{ols}} = \mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j(y - \boldsymbol{Z}_j\boldsymbol{\delta})^2\Big)\mathbb{E}\Big(\boldsymbol{Z}_j^T\boldsymbol{Z}_j\Big)^{-1}$ is the asymptotic variance matrix of the OLS estimator $\hat{\boldsymbol{\delta}} = [\hat{\boldsymbol{\beta}}^T,\ \hat{\boldsymbol{\gamma}}^T]^T$ from Step 1.

The proof follows from writing the two-step estimator in the GMM framework (cf. Newey and McFadden 1994); see Appendix C.1 for details.

preferred). In that sense, external conditions influence choices only to the extent they change internal psychological motivations.

For the sake of precision, suppose we are concerned with the choice of $y \in \mathcal{Y}$ (such as the amount purchased of a given item) in a variety of settings j (such as items within a category) and treatment states w (such as price). Each such decision problem induces a menu of motivational attribute bundles, $\{\boldsymbol{\theta}_j(y,w)\}_{y\in\mathcal{Y}}$ (where $\boldsymbol{\theta}_j(y,w)$ is the bundle for option y). If there are two settings j and j' along with treatments w_j and $w_{j'}$ for which $\boldsymbol{\theta}_j(y,w_j) = \boldsymbol{\theta}_{j'}(y,w_{j'})$ for all $y \in \mathcal{Y}$, then our premise is that a person would choose the same value of y in either.[26]

If we replaced $\boldsymbol{H}_j(w)$ with variables $\boldsymbol{\Theta}_j(w)$ that govern the relationship between y and $\boldsymbol{\theta}_j(y,w)$,[27] Assumption 1 would simply restate our central premise that decisions depend only on internal psychological motivations for the problem at hand.[28]

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

Consideration of this motivating framework suggests that our approach is most suitable in applications that have the following features.

Outcomes are choices of individuals. Our methods rely on respondents to evaluate factors that predict outcomes. This procedure makes sense when the outcomes are individual choices, but not necessarily when they result from technological or biological processes that respondents may poorly understand. For example, if the objective were to measure the effect of water purification on health, one could ask community members to predict health outcomes with and without this intervention, and extrapolate based on the observed relationship between actual health outcomes and predicted outcomes that vary for other reasons. To understand why Assumption 1 might fail in this context, suppose the observed variation in outcomes and evaluations is associated with a condition (other than the

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

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

[28] Technically, stochastic variation would be de minimis, and conditioning on $\boldsymbol{X}$ would be unnecessary.

treatment) that respondents understand much better than they understand the treatment. In that case, the relationship between health and hypothetical evaluations will not be the same for within-sample variation and treatment variation. Our approach will then produce biased estimates of the treatment's effects.

Similar issues arise when the outcome results from a collection of interacting decisions—in particular, when it is a feature of an equilibrium rather than a choice considered in isolation. For example, the effect of the minimum wage on equilibrium employment depends on interactions between the decisions of employers and prospective employees. Asking respondents to predict an equilibrium outcome is much like asking them to predict the outcome of a technological or biological process. However, one could use our method to analyze partial equilibrium effects of a minimum wage on job search by prospective employees, and, separately, on employers' hiring practices.

Hypothetical evaluations adequately proxy for motivations. The adequacy with which hypothetical evaluations proxy for motivations, and consequently the plausibility of Assumption 1, depends on a number of considerations, some of which the analyst controls.

Evaluators are similar to decision makers. Evaluators who more closely resemble the decision makers are likely to have better information about their motivations. That consideration argues for sampling respondents from the population that makes choices, with minimal temporal separation. But doing so introduces the possibility that respondents' real choices distort their hypothetical evaluations (for example, through anchoring or ex post rationalization). This possibility is less of a concern for decisions that are differentiated or forgettable. For example, our microlending respondents probably did not make decisions, or recall making decisions, about the particular profiles they evaluated hypothetically. In some applications, it may be possible to mitigate the concern by eliciting hypothetical evaluations prior to the treatment's implementation, or by identifying a similar but unexposed subpopulation.

Choice scenarios are familiar or natural. Hypothetical evaluations are more likely to be informative when descriptions of the choice scenarios bring all the relevant information to mind. In some applications, these scenarios may be so standard that a short hypothetical description suffices (as in our first application, which involves purchases of common snack foods). In others, it may be possible to depict the choice scenarios naturalistically (as in our second application, which involves online microfinance lending). Our method is less likely to work when the study examines choices that are unfamiliar or too complex to fully

represent when gathering hypothetical evaluations (for example, hypothetical automobile purchase decisions cannot include test drives).

Evaluations span motivational attributes. The set of hypothetical evaluations should be rich enough to span the factors underlying the available motivational attribute bundles. This spanning can be attained by combining broad composites (such as hypothetical choices, which summarize a collection of attributes) with narrowly focused evaluations (such as the intensity of temptation an option evokes). The use of composites, broad and narrow, avoids the need to fully catalog motivational attributes, and to pair each with a matching proxy.[29]

To understand the logic of the spanning requirement, imagine that the hypothetical evaluations $\boldsymbol{H}$ are each linear functions of $\boldsymbol{\Theta}$. In that case, any linear function of $\boldsymbol{H}$ is implicitly a linear function of $\boldsymbol{\Theta}$. Now imagine that y is also a linear function of $\boldsymbol{\Theta}$. The purpose of the spanning condition is to ensure that, by appropriately reweighting the elements of $\boldsymbol{H}$ (as a regression would do), we can reproduce any linear function of $\boldsymbol{\Theta}$, including the one that describes y.

Even when hypothetical evaluations span the underlying space of motivational attributes, the empirical relationship between $\boldsymbol{\Theta}$ and $\boldsymbol{H}$ (and hence between y and $\boldsymbol{H}$), may be unstable, contrary to Assumption 1. One potential reason for instability is that respondents may report $\boldsymbol{H}$ with hypothetical biases that vary across settings. We can address this source of instability by expanding $\boldsymbol{H}$ so that it also spans the motivations that impact reporting biases, such as the extent to which others would approve of each response.[30] Second, the relationship between $\boldsymbol{H}$ and $\boldsymbol{\Theta}$ may depend on extraneous factors, such as measurement error, which we address in the next subsection.

5.3 Diagnostic checks and extensions

We can relax the linearity assumption when overlap between the marginal distributions of $\boldsymbol{H}_j(W_j)$ and $\boldsymbol{H}_j(1-W_j)$ is high (see Appendix D.5). Treatment effects are then identified semiparametrically, and one can dispense with functional form assumptions entirely; see

[29] One does not actually need $\boldsymbol{H}$ to subsume *all* the information contained in $\boldsymbol{\Theta}$: a natural possibility is that people answer hypothetical questions by envisioning *typical* decision conditions, rather than the *specific* conditions that give rise to the observed value of y and the associated latent value of $\boldsymbol{\Theta}$. As long as the idiosyncratic effects of these specific conditions are orthogonal to the information contained in $\boldsymbol{H}$ as well as to the treatment W, this consideration simply adds randomness to the relationship between y and $\boldsymbol{H}$ without overturning Assumption 1. See Appendix F for an elaboration of this possibility.

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

Appendix C.2, which replaces Assumption 2 (Linearity) with a different Assumption 4 (Evaluations overlap).

As with standard methods, results that are robust across progressively richer specifications may instill greater confidence. As one adds evaluations, the risk of overfitting grows, particularly if the specifications include transformations and interactions to provide functional form flexibility. In that case, one may also use machine learning to select among the potential predictors; see Appendices C.3 and C.4.

When survey samples are small and not easily expanded, sampling error in $\boldsymbol{H}_j(0)$ and $\boldsymbol{H}_j(1)$ can potentially bias our estimator. One can address this concern by employing standard corrections for measurement error. Appendix C.3.3 illustrates one such correction.

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

6 Endogeneity

So far, our formal results show that the estimator recovers treatment effects if treatment has no variation (Assumption 3A) or is assigned randomly (Assumption 3B). This section considers how our approach performs when treatment is assigned endogenously.[31]

6.1 An assumption that limits endogeneity bias

In the absence of quasiexperimental variation, standard estimators infer the effect of treatment from differences between treated and untreated outcomes. The simplest standard estimator is $\hat{\tau} = \frac{1}{J}\sum_j \left[\frac{W_j}{p}Y_j(1) - \frac{1-W_j}{1-p}Y_j(0)\right]$ where $p = \frac{1}{J}\sum_j W_j$. Any systematic differ-

[31] Our approach models the choice of outcome. An alternative would be to model the choice of treatment by eliciting hypothetical evaluations from people resembling the treatment selectors (for example Briggs et al. 2020). This procedure may facilitate analyses of treatments affecting outcomes that are not choices. However, it may be difficult to survey people resembling those selecting the treatment (who may be specialists such as retail price strategists or sponsors of matches for charitable contributions).

ences between the treated and untreated observations, other than the treatment itself, directly confound $\hat{\tau}$.

In contrast, for our method, the same considerations introduce bias less directly. As explained in Section 2, treatment endogeneity can only bias the estimate by distorting the value of $\boldsymbol{\beta}$ obtained in Step 1. Because "ambient" variation in hypothetical evaluations helps identify the relationship between outcomes and hypotheticals, it can dilute any endogeneity problem. When this ambient variation is sufficiently important relative to the treatment, the bias becomes arbitrarily small. Thus, our method can identify causal effects accurately so long as most variation in hypotheticals is not from treatment. This condition is satisfied if treatment is rare, or when the sample is *diverse* in the sense that hypotheticals vary significantly for other reasons.

To formalize this intuition, we maintain the assumptions of mapping invariance and linearity, and focus on models that employ a single hypothetical evaluation without controlling for fixed characteristics. We then consider sequences of environments, indexed by k, such that as k grows the importance of the treatment relative to ambient variation in hypotheticals shrinks:

Assumption (3C). **Most variation not due to treatment.** *For $\tilde{w} = 0$ or $\tilde{w} = 1$: as $k \to \infty$, (i) $\frac{\text{var}_k(\mathbf{1}\{W_j=\tilde{w}\}(H_j(1)-H_j(0)))}{\text{var}_k(H_j(\tilde{w}))} \to 0$, (ii) $\frac{\text{var}_k(\mathbf{1}\{W_j=\tilde{w}\}(\epsilon_j(1)-\epsilon_j(0)))}{\text{var}_k(H_j(\tilde{w}))} \to 0$, and (iii) $\frac{\text{var}_k(\epsilon_j(\tilde{w}))}{\text{var}_k(H_j(\tilde{w}))} < \infty$.*

The conditions state that the "ambient" variation in hypotheticals is large relative to the variation in outcomes from the impact of the treatment through (i) the hypotheticals and (ii) the error terms. Condition (iii) requires that the model linking outcomes to hypotheticals does not become arbitrarily poor in terms of fit (e.g., that the R^2 in a regression of $Y_j(0)$ on $H_j(0)$ is bounded away from 0).

Adding this assumption yields the following result:

Proposition 2. *Suppose Assumptions 1, 2, and 3C hold, $0 < \text{var}_k(H_j) < \infty$ and $|\text{cov}(H_j, Y_j)| < \infty$ for all fixed k, and $\mathbb{E}_k(H_j(1) - H_j(0))$ is bounded. Then the asymptotic bias of $\hat{\tau}$ vanishes, i.e., $\lim_{k\to\infty} \text{plim}_{n\to\infty} \hat{\tau}_k - \tau_k = 0$.*[32]

The proposition above potentially applies whenever there is substantial natural variation in hypotheticals across settings, relative to the variation arising from treatment, so long as the hypotheticals are sufficiently informative about treatment effects and outcomes. As

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

a special case, the key conditions (i) and (ii) are satisfied if treatment is rare (probability of treatment $p_k := \mathbb{E}_k(W_j) \to 0$ as $k \to \infty$) or very common ($p_k \to 1$ as $k \to \infty$), provided potential outcomes and hypotheticals are bounded and have non-zero variance. The proposition also suggests that one can potentially reduce endogeneity bias by collecting choice data from more diverse settings. For example, in our snack application, one could create more ambient variation by oversampling snacks that are unusually desirable or undesirable according to the various subjective measures. However, there is a qualification: the model relating outcomes to hypotheticals must continue to perform well across the diverse settings. Combining different types of choices, such as purchases of snacks and automobiles, would make the settings more diverse but would likely undermine the model's predictive performance, which could violate condition (iii).

In contrast, increasing the diversity of observational settings does not reduce endogeneity bias for the standard estimator discussed above. Consider regressing the outcome on covariates, separately among the treated and the control. Increasing the variance of the covariates $\boldsymbol{X}_j$ may increase the precision with which one estimates $\mathbb{E}(Y_j(1) \mid W_j = 1, \boldsymbol{X}_j) = \boldsymbol{X}_j\boldsymbol{\beta}_1$ and $\mathbb{E}(Y_j(0) \mid W_j = 0, \boldsymbol{X}_j) = \boldsymbol{X}_j\boldsymbol{\beta}_0$ individually, but there is no reason to think it would decrease the difference between $\mathbb{E}(Y_j(1) \mid W_j = 1, \boldsymbol{X}_j)$ and $\mathbb{E}(Y_j(1) \mid \boldsymbol{X}_j)$, or between $\mathbb{E}(Y_j(0) \mid W_j = 0, \boldsymbol{X}_j)$ and $\mathbb{E}(Y_j(0) \mid \boldsymbol{X}_j)$. Additionally, if treatment became more rare, the main effect would be to reduce the precision with which we estimate $\mathbb{E}(Y_j(1) \mid W_j = 1, \boldsymbol{X}_j)$, without necessarily reducing the bias. Furthermore, in the limit of $p = 0$, the standard estimator becomes undefined.

One can assess the conditions in Assumption 3C informally by computing suggestive diagnostics. Condition (i) requires the variance of $W_j(H_j(1) - H_j(0))$ (treatment times treatment effect) to be small relative to the variance of $H_j(0)$ (hypotheticals). In the microfinance application, the ratio $\frac{\text{var}(W_j(H_j(1)-H_j(0))}{\text{var}(H_j(0))}$ ranges from 0.041 to 0.074 for the regressors constructed from hypothetical evaluations used in Column (7) of Table 3, suggesting that treatment accounts for a small share of the variation in hypotheticals. Condition (iii) in Assumption 3C requires that hypotheticals contribute meaningfully to the variation in choices. In the microfinance application, a regression of $Y_j(0)$ on $\boldsymbol{H}_j(0)$ and $\boldsymbol{X}_j$ among the untreated observation yields $R^2 = 0.21$, which suggests that hypotheticals have decent explanatory power.[33]

[33] One would ideally assess condition (iii) by examining the R^2 of a regression of $Y_j(0)$ on $H_j(0)$ using all observations (treated and control). That regression is infeasible, but a similar regression using observations in a single treatment state is feasible.

6.2 Deriving bounds that are robust to treatment assignment

Our main theoretical results describe the properties of point estimates under Assumptions 1 and 2 combined with one assumption governing treatment assignment: either Assumption 3A, 3B, 3C, or 3D (which we will introduce in the next section). However, even without one of the latter assumptions, one can still recover useful bounds on treatment effects that apply regardless of any endogeneity.

We begin by describing the logic of our bounding procedure informally. Let $\tau_j \equiv Y_j(1) - Y_j(0)$ be the treatment effect for observation j. How might we rule out the possibility that a particular vector, $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, ..., \tilde{\tau}_J)$, equals the vector of treatment effects, i.e., $\tilde{\tau}_j = \tau_j$, using only linearity and mapping invariance? If a given $\tilde{\boldsymbol{\tau}} \in \mathbb{R}^J$ were the true treatment effect vector, then we could transform the observational sample into a synthetic sample containing only treated observations: keep $Y_j(1)$ for treated observations and construct $\tilde{Y}_j(1) = Y_j + \tilde{\tau}_j$ for untreated observations. If $\tilde{\tau}_j = \tau_j$ then, since $Y_j = Y_j(0)$, we would have $\tilde{Y}_j(1) = Y_j(1)$, which means a regression of $\tilde{\boldsymbol{Y}}(1)$ on $\boldsymbol{H}(1)$ and $\boldsymbol{X}$ using the synthetic sample would yield consistent estimates of the relationship between potential outcomes, hypotheticals, and other factors. Similarly, we could transform the original sample into a second synthetic sample with no treatment by constructing $\tilde{Y}_j(0) = Y_j - \tilde{\tau}_j$ for treated observations and keeping $Y_j(0)$ for untreated observations. A regression of $\tilde{\boldsymbol{Y}}(0)$ on $\boldsymbol{H}(0)$ and $\boldsymbol{X}$ using the second synthetic sample would likewise yield consistent estimates of the relationship between potential outcomes, hypotheticals, and other factors. Furthermore, under our mapping invariance assumption, these two regression equations are the same. So if the implied regression equations do not coincide, $\tilde{\boldsymbol{\tau}}$ cannot be the true vector of treatment effects, $\tilde{\boldsymbol{\tau}} \neq \boldsymbol{\tau}$. We obtain bounds by ruling out values of the average treatment effect that are not decomposable into treatment effect vectors satisfying the requirement that the regression equations are the same. The next proposition formalizes this intuition.

Proposition 3. *Suppose Assumptions 1 and 2 hold in the sample.*[34] *Then, for any user-specified bounds (including no bounds) on potential outcomes* $\underline{Y}, \bar{Y} \in \mathbb{R} \cup \{-\infty, \infty\}$ *and setting-specific treatment effects* $\underline{\tau}, \bar{\tau} \in \mathbb{R} \cup \{-\infty, \infty\}$*, the average in-sample treatment effect is bounded by:*

[34] Specifically, suppose that $(\boldsymbol{Z}(0)'\boldsymbol{Z}(0))^{-1}\boldsymbol{Z}(0)'\boldsymbol{Y}(0) = (\boldsymbol{Z}(1)'\boldsymbol{Z}(1))^{-1}\boldsymbol{Z}(1)'\boldsymbol{Y}(1)$ for the observations in the sample. This assumption is in the same spirit as assuming, for OLS, that $\mathbb{E}(\boldsymbol{X}_j^T(Y_j - \boldsymbol{X}_j\boldsymbol{\beta})) = 0$ and then estimating $\boldsymbol{\beta}$ by imposing the sample analog $\frac{1}{J}\sum_{j=1}^{J}\boldsymbol{X}_j^T(Y_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}) = 0$, yielding $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ for regressor matrix $\boldsymbol{X}$ and outcome vector $\boldsymbol{Y}$.

$$\tau_{lb} = \min_{\tilde{\boldsymbol{\tau}} \in \mathbb{R}^J} \quad \frac{1}{J}\sum_{j=1}^{J} \tilde{\tau}_j \qquad\qquad \tau_{ub} = \max_{\tilde{\boldsymbol{\tau}} \in \mathbb{R}^J} \quad \frac{1}{J}\sum_{j=1}^{J} \tilde{\tau}_j$$

$$\textit{subject to:} \quad \tilde{\boldsymbol{\tau}} \in \mathbb{C}(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}(0), \boldsymbol{Z}(1)) \qquad \textit{subject to:} \quad \tilde{\boldsymbol{\tau}} \in \mathbb{C}(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}(0), \boldsymbol{Z}(1))$$

for constraints given by

$$\begin{aligned}\mathbb{C}(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}(0), \boldsymbol{Z}(1)) = \quad \{\tilde{\boldsymbol{\tau}} \in \mathbb{R}^J : & (\boldsymbol{Z}(0)'\boldsymbol{Z}(0))^{-1}\boldsymbol{Z}(0)'(\boldsymbol{Y} - \boldsymbol{W}\tilde{\boldsymbol{\tau}}) = (\boldsymbol{Z}(1)'\boldsymbol{Z}(1))^{-1}\boldsymbol{Z}(1)'(\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{W})\tilde{\boldsymbol{\tau}}) \\ & \underline{Y} \leq Y_j - W_j\tilde{\tau}_j + (1 - W_j)\tilde{\tau}_j \leq \bar{Y}, \\ & \underline{\tau} \leq \tilde{\tau}_j \leq \bar{\tau}, \textit{ for } j = 1, \ldots, J\}\end{aligned}$$

where $\boldsymbol{Z}(0)$ and $\boldsymbol{Z}(1)$ are the matrices of regressors $\boldsymbol{Z}_j(w) = [\boldsymbol{H}_j(w),\ \boldsymbol{X}_j]$ (hypotheticals and fixed characteristics including the intercept) if all observations (i.e. having J rows) were untreated and treated, respectively, $\boldsymbol{Y} = (Y_1, \ldots, Y_j)^T$, $\boldsymbol{I}$ is the $J \times J$ identity matrix, $\boldsymbol{W}$ is the $J \times J$ diagonal matrix with (j, j) diagonal element equal to W_j, and $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_J)^T$.

In each case, the objective function is the average treatment effect if the (unknown) effect of the treatment on setting j is $\tau_j = \tilde{\tau}_j$. The first constraint imposes mapping invariance and linearity, assuming that the elements of $\tilde{\boldsymbol{\tau}}$ are the setting-specific treatment effects: $(\boldsymbol{Y} - \boldsymbol{W}\tilde{\boldsymbol{\tau}}) = \boldsymbol{Y}(0)$ and $(\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{W})\tilde{\boldsymbol{\tau}}) = \boldsymbol{Y}(1)$ when $\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau}$. (Appendix C.6.1 derives bounds without assuming linearity.) The remaining constraints can tighten the resulting bounds on average treatment effects if one is willing to limit the range of potential outcomes ($\underline{Y}$, $\bar{Y}$) or treatment effects ($\underline{\tau}$, $\bar{\tau}$). The problem is a linear program (in $\tilde{\boldsymbol{\tau}}$), so one can compute its solution efficiently.

In the microfinance application, without any constraints, the average treatment effect must lie between $(-\infty, \infty)$. One could impose only "natural" bounds on potential outcomes in the spirit of Manski (1990), that funding rates must be non-negative ($\underline{Y} = 0$) and the largest loans are never fully funded in less than one minute ($\bar{Y} = 14.87$).[35] However, those constraints alone yield wide bounds on the average treatment effect: $[-5.02, 9.85]$. Adding mapping invariance and linearity (using the specification in Column (8) of Table 3) substantially narrows these bounds to $[0.37, 1.88]$, even without constraining treatment effects ($\underline{\tau} = -\infty, \bar{\tau} = \infty$). Adding a constraint that matching cannot *reduce* funding for any loan ($\underline{\tau} = 0$, $\bar{\tau} = \infty$) shrinks the bounds further, to $[0.67, 1.25]$.[36] Altogether,

[35] Among observed outcomes, $\max_j Y_j = 11.87$, equivalent to the largest loans being funded in 20 minutes.

[36] Note that the point estimate based on the same assumptions can be outside the bounds due to sampling uncertainty. While it is challenging to compute confidence intervals in partially identified models based on

mapping invariance and linearity have much identifying power, even without assumptions on treatment assignment.

6.3 Alternate assumptions for treatment assignment

Two alternative assumptions yield consistency and asymptotic normality.

Unconfoundedness Our basic estimator is consistent, following Proposition 1, under an analog to the standard unconfoundedness assumption:

Assumption (3D)**. Unconfoundedness.** *Treatment assignment is unconfounded conditional on hypothetical evaluations*: for $w \in \{0, 1\}$, $W_j \perp\!\!\!\perp Y_j(w) \mid \boldsymbol{H}_j(w), \boldsymbol{X}_j$.

For unconfoundedness to hold, the control variables would have to span the outcome-relevant information used to select the treatment. Hypothetical evaluations may be useful controls because they may resemble that information. For example, legislators may have implemented a treatment partly based on public opinion polls years before the earliest observation in the dataset, and outcome-relevant information at the time of passage may be limited to broad attitudinal considerations that later surveys can easily capture.

An assumption related to unconfoundedness, $W_j \perp\!\!\!\perp Y_j(0), Y_j(1) \mid \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j$, justifies running a standard regression of outcomes on a treatment dummy "controlling for" $\boldsymbol{H}_j(0)$, $\boldsymbol{H}_j(1)$, and $\boldsymbol{X}_j$. If those estimates are similar to the ones derived from our method, the unconfoundedness assumption may be plausible. We find that the two approaches yield similar estimates in the snack application but not the microfinance application. This finding suggests that, for our microfinance application, unconfoundedness may not be plausible. The accuracy of our method in that setting is likely attributable to the considerations formalized in Propositions 2 and 3.

Sample selection model A regression of $Y_j(w_j)$ on $\boldsymbol{H}_j(w_j)$, and $\boldsymbol{X}_j$ for the subset of observations with $W_j = 0$, or for the subset with $W_j = 1$, would yield consistent estimates of the underlying relationship between outcomes and hypothetical evaluations if treatment assignment, and hence the selection of these subsets, were random. Treatment endogeneity potentially biases the coefficient estimates for these regressions, and hence for our method, because it makes the selection of the $w = 0$ and $w = 1$ samples non-random. Accordingly, in

linear programs where the number of constraints grows with the sample size, we describe an approximation in Appendix C.6.2.

our setting, one can address treatment endogeneity through modified versions of standard sample selection corrections.

In Appendix C.7, we model treatment assignment using an approach similar to that of Heckman (1976). An important feature of this approach is that the hypothetical evaluation for the unobserved treatment state serves as an internal instrument, so that identification does not rely exclusively on the functional form. In the microfinance application, the estimate of the ATC is relatively insensitive to the particular assumption on treatment assignment used to obtain point identification. Using the same specification as in Column (8) of Table 3, but correcting for sample selection, we obtain an estimate of 1.31 (bootstrap s.e. 0.46).

7 Conclusion

We have proposed estimators that infer the causal effects of treatments on choices by combining real choices and hypothetical evaluations based on a new "mapping invariance" assumption. We have explored the implications of this method theoretically and demonstrated that our estimators yield promising results in both a laboratory application and a field application. The approach is not a panacea, but adds an additional tool to the causal inference toolbox for suitable applications. An important question is whether the relationship between choices and basic motivations is stable, and therefore portable, over a broad domain. 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 be possible to develop more universal mappings between hypothetical responses and real choices.

References

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

Abdellaoui, Mohammed, Carolina Barrios, and Peter P. Wakker (2007). "Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory". en. In: *Journal of Econometrics*. 50th Anniversary Econometric Institute 138.1, pp. 356–378.

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

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

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

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

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

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

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

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

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

Blumenschein, Karen, Glenn C. Blomquist, Magnus Johannesson, Nancy Horn, and Patricia Freeman (2008). "Eliciting Willingness to Pay Without Bias: Evidence from a Field Experiment". In: *The Economic Journal* 118.525, pp. 114–137.

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

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

Brownstone, David, David S. Bunch, and Kenneth Train (2000). "Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles". en. In: *Transportation Research Part B: Methodological* 34.5, pp. 315–338.

Carson, Richard T. (2012). "Contingent Valuation: A Practical Alternative When Prices Aren't Available". en. In: *Journal of Economic Perspectives* 26.4, pp. 27–42.

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

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

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". en. In: *The Econometrics Journal* 21.1, pp. C1–C68.

Cummings, Ronald G., Glenn W. Harrison, and E. Elisabet Rutström (1995). "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible?" In: *The American Economic Review* 85.1, pp. 260–266.

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

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

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

Gruber, Jonathan and Ebonya Washington (2005). "Subsidies to employee health insurance premiums and the health insurance market". In: *Journal of Health Economics* 24.2, pp. 253–276.

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

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

Imbens, Guido W. and Joshua D. Angrist (1994). “Identification and Estimation of Local Average Treatment Effects”. In: *Econometrica* 62.2, pp. 467–475.

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

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

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

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

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

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

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

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

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

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

Levy, Ifat, Stephanie C. Lazzaro, Robb B. Rutledge, and Paul W. Glimcher (2011). “Choice from Non-Choice: Predicting Consumer Preferences from Blood Oxygenation Level-

Dependent Signals Obtained during Passive Viewing". en. In: *Journal of Neuroscience* 31.1, pp. 118–125.

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

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

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

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

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

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

Magnolfi, Lorenzo, Jonathon McClure, and Alan T. Sorensen (2022). *Triplet Embeddings for Demand Estimation*. en. SSRN Scholarly Paper 4113399. Rochester, NY: Social Science Research Network.

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

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

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

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

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

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

Rosenbaum, Paul R. and Donald B. Rubin (1983). “The central role of the propensity score in observational studies for causal effects”. en. In: *Biometrika* 70.1, pp. 41–55.

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

Rothschild, David M. and Justin Wolfers (2011a). “Forecasting Elections: Voter Intentions Versus Expectations”. In: *mimeo*.

— (2011b). *Forecasting Elections: Voter Intentions Versus Expectations*. en. SSRN Scholarly Paper ID 1884644. Rochester, NY: Social Science Research Network.

Shenhav, Amitai (2024). “The affective gradient hypothesis: an affect-centered account of motivated behavior”. English. In: *Trends in Cognitive Sciences* 28.12, pp. 1089–1104.

Shogren, Jason F. (2006). “Valuation in the lab”. In: *Environmental and resource Economics* 34.1, pp. 163–172.

Small, Kenneth A., Clifford Winston, and Jia Yan (2005). “Uncovering the Distribution of Motorists’ Preferences for Travel Time and Reliability”. en. In: *Econometrica* 73.4, pp. 1367–1382.

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

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

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

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

Online Appendices

A Appendix Figures and Tables

Figure A1: Snack Demand Application: A Typical Choice Task

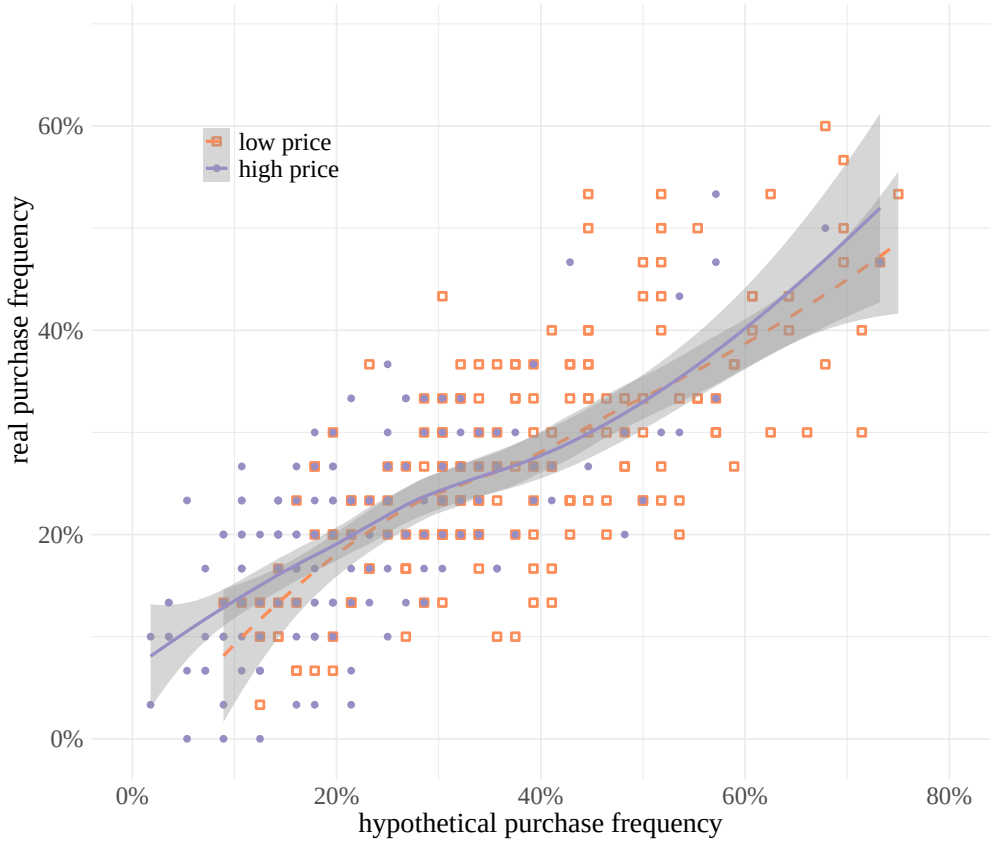
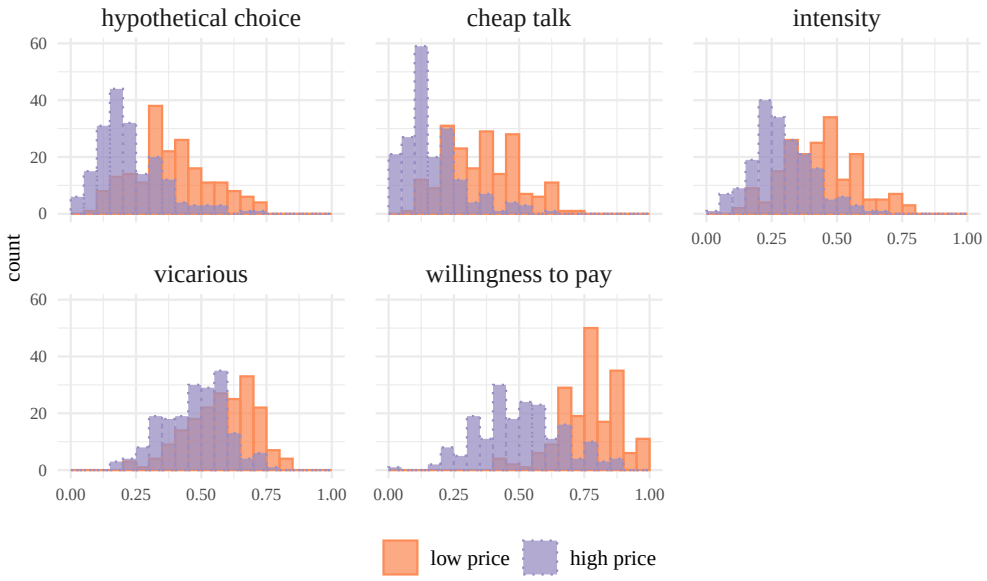
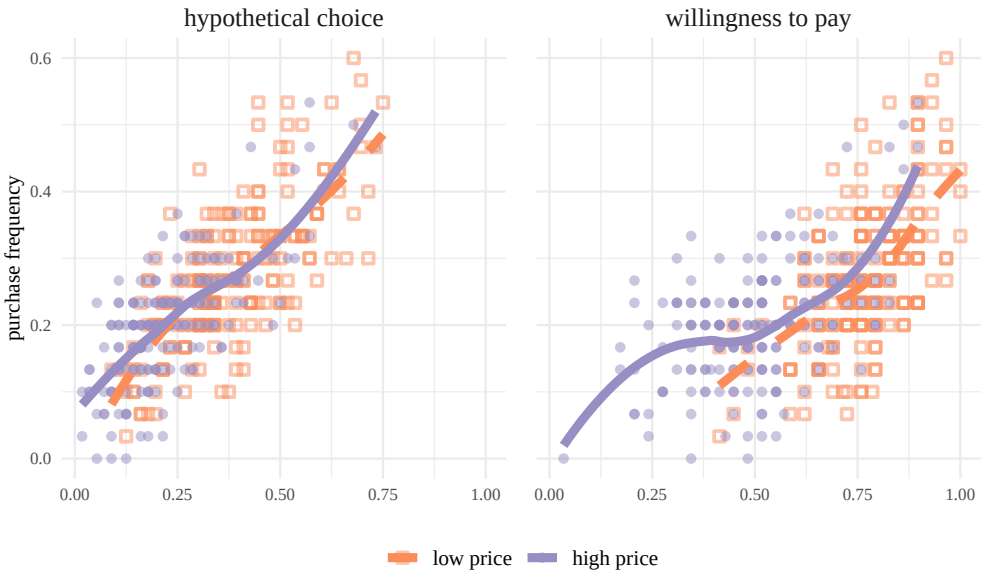


Figure A2: Real vs. Hypothetical Choices (nonparametric) in Snack Demand Application

Item-price pairs plotted. The lines and pointwise 95% confidence intervals shown are based on local quadratic regressions using the `R` function `stats::loess` with default parameters: Estimates at each point use 75% of the data nearest to the point with tricubic weighting.



(a) Overlap in Hypotheticals



(b) Relationship between Outcome and Hypothetical

Figure A3: Overlap between Hypothetical Evaluations in Snack Demand Application

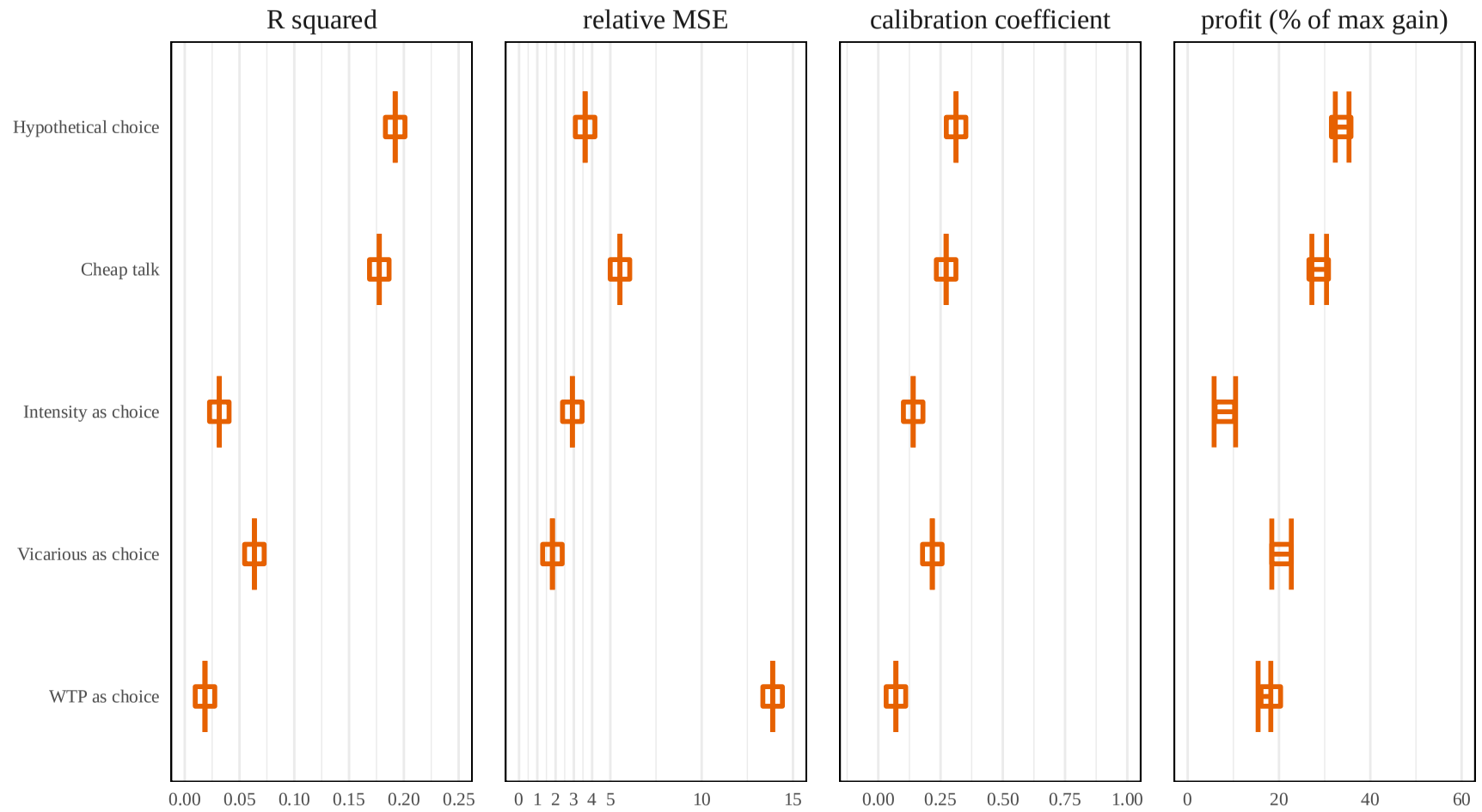


Figure A4: Treatment Effect Heterogeneity Using Solely Hypotheticals As Predictions in Snack Demand Application

Summary statistics describing how well different standard estimators based on a difference in hypothetical choices capture heterogeneity in treatment effects. Each row shows the results for hypothetical choices elicited with a different protocol, as described in Section 3.1.2. Points indicate the median value of each statistic across 10,001 simulated samples, and whiskers indicate the interquartile range. 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.

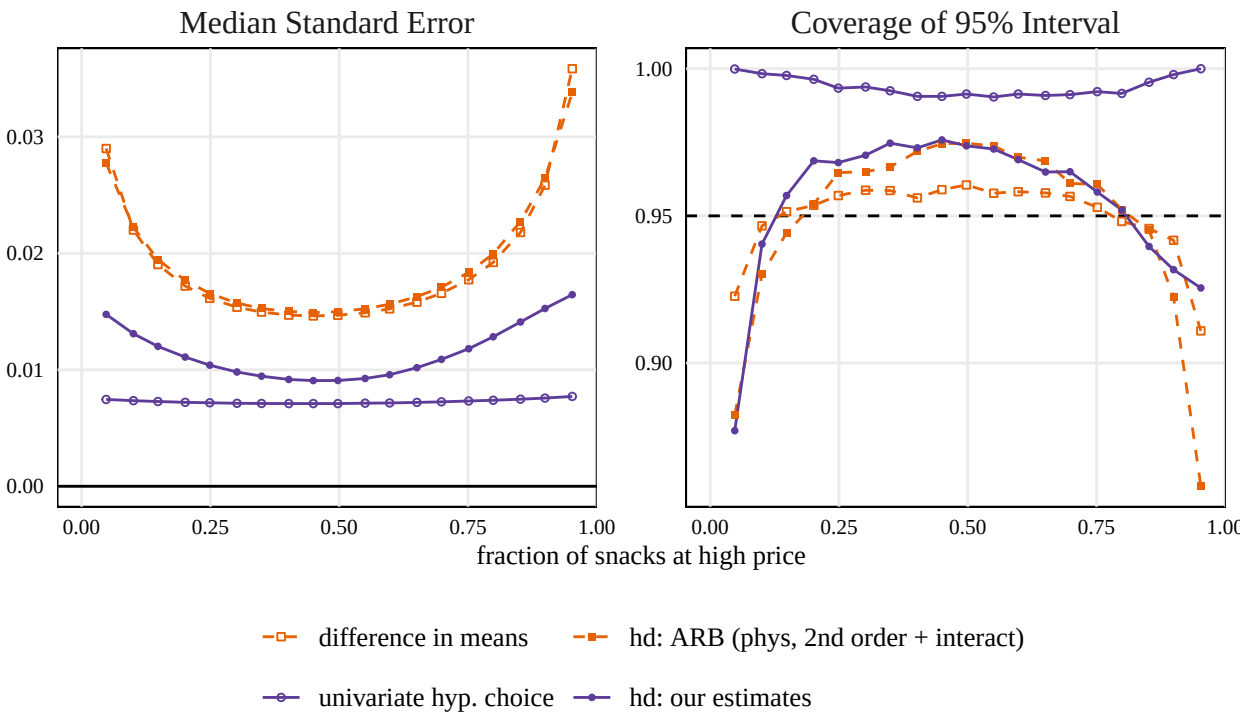


Figure A5: Performance of Estimators by Fraction Treated in Snack Demand Application

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

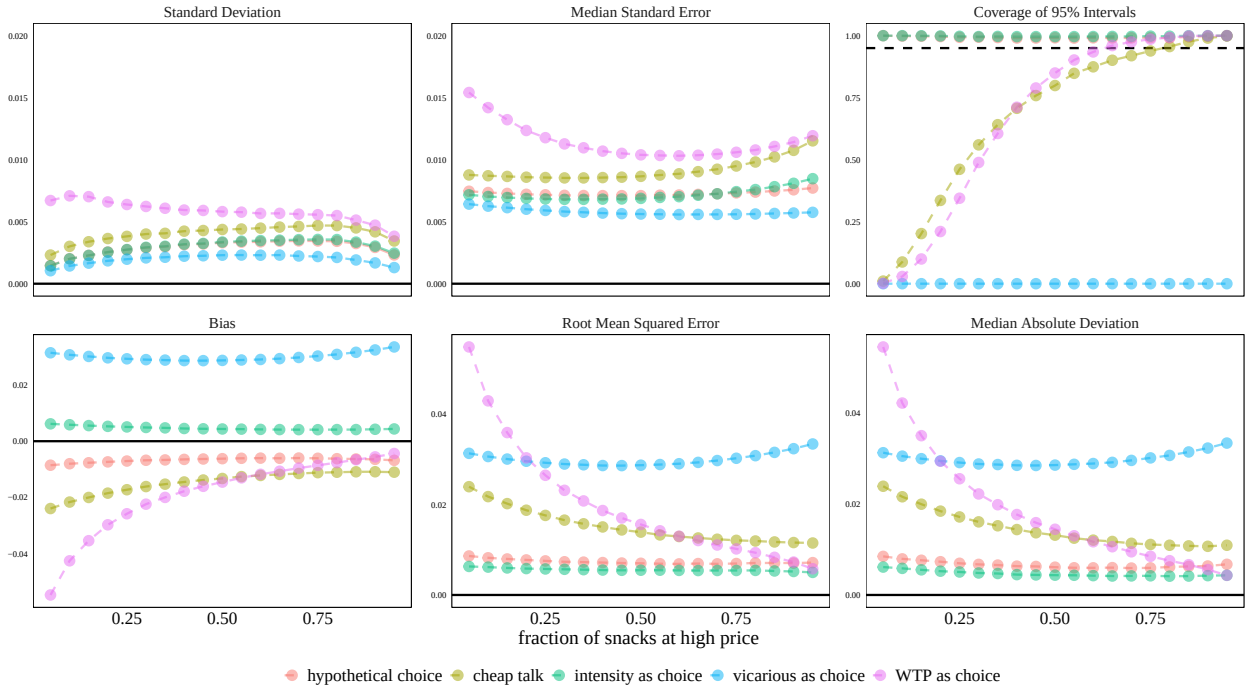


Figure A6: Performance of Univariate Estimators by Fraction Treated in Snack Demand Application

Summary statistics describing properties of univariate treatment effect estimators under random assignment with 10,001 samples varying only treatment assignment for each fraction of snacks treated. The horizontal axis measures the fraction of snacks observed at the high price. The estimators correspond to columns (2) through (6) in panel (b) of Table 1 in the main text. Coverage of confidence intervals refers to coverage of the true in-sample treatment effect.

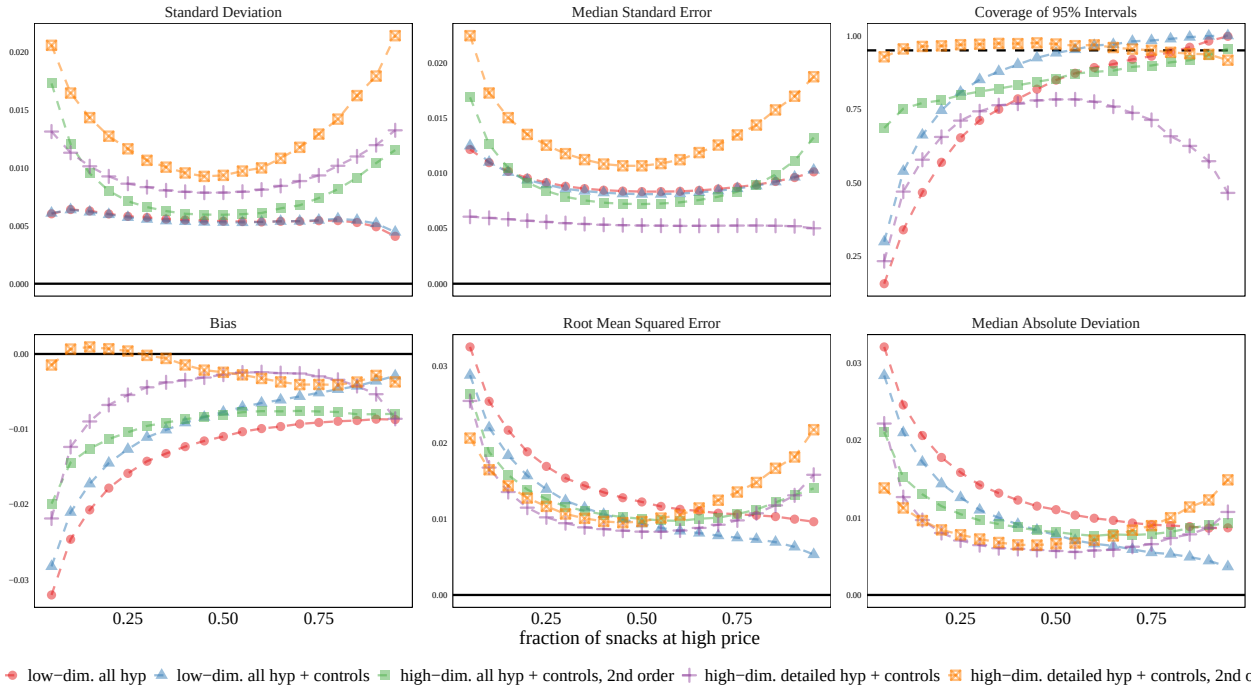


Figure A7: Performance of Multivariate Estimators by Fraction Treated in Snack Demand Application

Summary statistics describing properties of multivariate treatment effect estimators under random assignment with 10,001 samples varying only treatment assignment for each fraction of snacks treated. The horizontal axis measures the fraction of snacks observed at the high price. The estimators correspond to columns (2) through (6) of Table 2 in the main text. Coverage of confidence intervals refers to coverage of the true in-sample treatment effect.

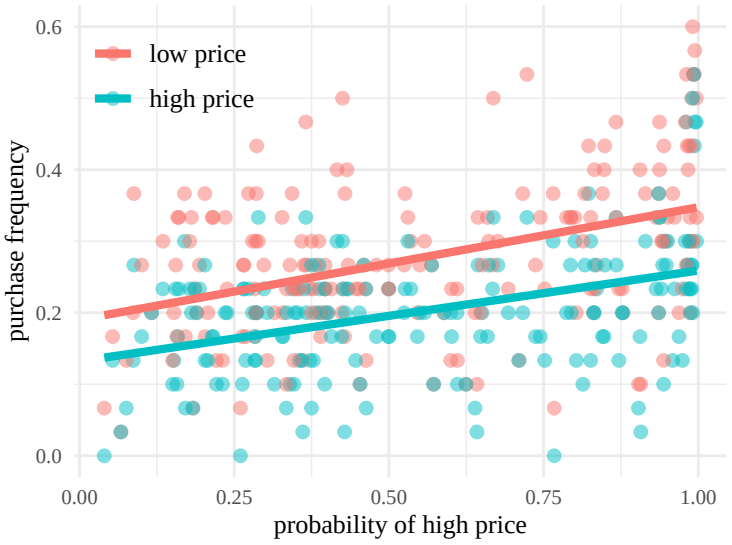


Figure A8: Endogenous Snack Pricing Simulation Setup

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.

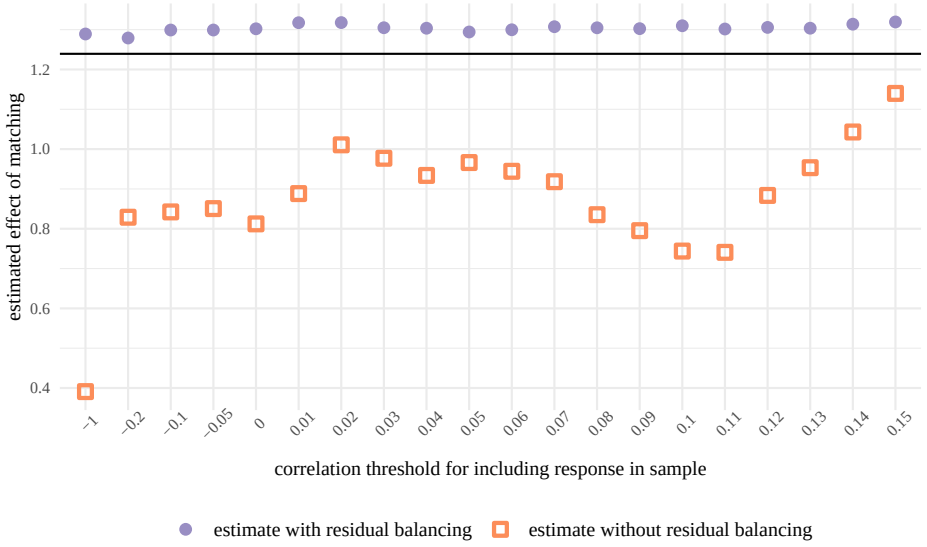
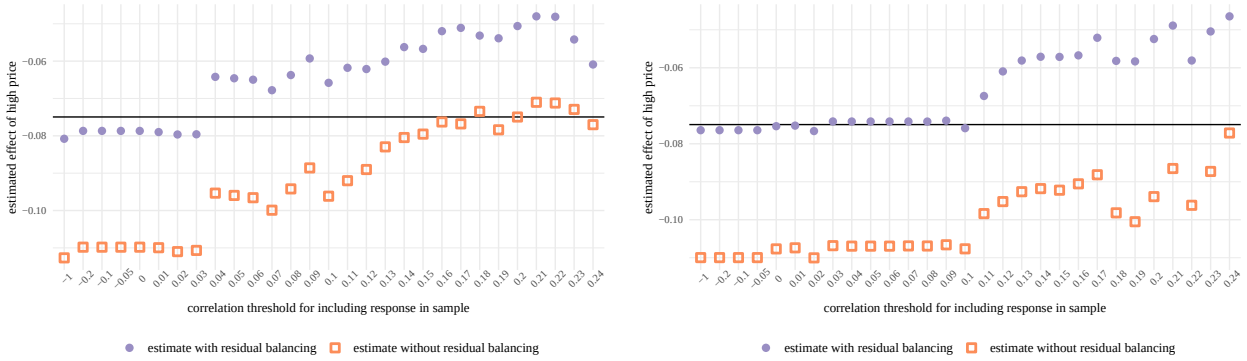


Figure A9: Estimates of the Effect of Matching by Correlation Threshold in Microfinance Application.

The horizontal line indicates the true in-sample treatment effect.



(a) Observing all snacks at high price

(b) Observing all snacks at low price

Figure A10: Estimates of the Effect of High Price by Correlation Threshold in Snack Demand Application

The horizontal line indicates the true in-sample treatment effect.

Table A1: Estimating Treatment Effects without Variation in Treatment in Snack Demand Application (multivariate, including WTP variable)

	Ground Truth	Our Method: Hypotheticals as Predictors				
		Low Dimensional		High Dimensional		
	(1)	(2)	(3)	(4)	(5)	(6)
Observing all snacks at high price						
Estimated effect of high price	-0.075	-0.084	-0.077	-0.085	-0.083	-0.094
	(0.004)	(0.011)	(0.011)	(0.016)	(0.005)	(0.014)
	[0.004]	[0.010]	[0.010]	[0.021]	[0.020]	[0.026]
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	-0.119	-0.116	-0.136	-0.122	-0.066
	(0.004)	(0.013)	(0.014)	(0.021)	(0.006)	(0.029)
	[0.004]	[0.013]	[0.013]	[0.020]	[0.025]	[0.028]
Observed at high price	All			None		
Observed at low price	All			All		
Controls			X	X	X	X
Hypotheticals:						
. . . all hypothetical choices (incl. WTP)		X	X	X	X	X
. . . detailed hypothetical eval. (incl. WTP)					X	X
2nd order + interactions				X		X
Sample size (outcome)	189 ($\times 2$)	189	189	189	189	189

Estimates of the effect of the high price (vs. low price) on the real purchase frequency. Analytical standard errors are in parentheses; bootstrap standard errors in square brackets are based on 1,001 bootstrap samples. For results excluding WTP see Table 2.

Table A2: Summary Statistics Describing Heterogeneous Treatment Effect Estimates in Snack Demand Application

	Benchmarks		Observational			Our Method				
	Diff. in Outcomes	Infeasible: OLS	OLS	LASSO	RF	Low-Dim.		High-Dim.		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Random Assignment										
median R squared	0	0.05	0.00	0.00	0.00	0.22	0.21	0.23	0.22	0.23
			(0.00–0.01)	(0.00–0.01)	(0.00–0.01)	(0.21–0.22)	(0.20–0.21)	(0.21–0.24)	(0.21–0.23)	(0.21–0.25)
median MSE (relative to ATE)	1.02	0.95	1.56	1.54	1.04	0.85	0.83	0.79	0.81	0.79
	(1.00–1.07)		(1.39–1.78)	(1.34–1.76)	(1.02–1.09)	(0.83–0.87)	(0.81–0.85)	(0.78–0.81)	(0.79–0.83)	(0.77–0.81)
median calibration coefficient	0	1	0.01	0.00	0.00	0.72	0.77	0.84	0.81	0.86
			(−0.06–0.08)	(−0.06–0.05)	(−0.32–0.31)	(0.69–0.75)	(0.74–0.81)	(0.80–0.88)	(0.76–0.86)	(0.80–0.91)
median profit (as % of max gain)	10	9	9	9	10	46	44	43	45	45
	(5–14)	(6–13)	(4–15)	(3–14)	(6–14)	(43–49)	(42–47)	(40–46)	(42–48)	(41–48)
Endogenous Assignment										
median R squared	0	0.05	0.00	0.00	0.00	0.19	0.18	0.18	0.19	0.20
			(0.00–0.01)	(0.00–0.01)	(0.00–0.01)	(0.18–0.20)	(0.16–0.19)	(0.15–0.20)	(0.18–0.20)	(0.18–0.22)
median MSE (relative to ATE)	1.66	0.95	2.48	2.39	1.57	0.86	0.85	0.85	0.83	0.83
	(1.49–1.85)		(2.19–2.77)	(2.10–2.77)	(1.42–1.74)	(0.84–0.87)	(0.84–0.87)	(0.82–0.88)	(0.81–0.85)	(0.80–0.87)
median calibration coefficient	0	1	0.00	−0.01	−0.06	0.70	0.73	0.76	0.79	0.83
			(−0.05–0.05)	(−0.07–0.04)	(−0.31–0.17)	(0.67–0.73)	(0.70–0.76)	(0.71–0.80)	(0.75–0.83)	(0.77–0.88)
median profit (as % of max gain)	−5	13	1	−1	−2	45	42	40	42	41
	(−8–−2)	11–16	(−4–7)	(−6–5)	(−6–3)	(42–47)	(38–44)	(37–42)	(39–45)	(37–45)
Controls		X	X	X	X		X	X	X	X
Hypotheticals:										
. . . all hypothetical choices						X	X	X	X	X
. . . detailed hypothetical eval.									X	X
2nd order + interactions				X				X		X

Median statistics (and interquartile range in parentheses) of different estimators across 10,001 samples, varying only treatment assignment, with random treatment assignment (top) and endogenous assignment as in Section 4.2 (bottom). Interquartile range is omitted when the respective statistic is mechanically 0 ("Diff. in Outcomes" does estimates no heterogeneity) and for the column using projected true effects ("Infeasible: OLS") because these do not vary across simulated samples, except for profit. Mean squared error (MSE) is shown relative to the MSE of the ATE as a prediction of setting-specific treatment effects. The table shows specifications used in Figure 3 as well as the other specifications of our method from Table 2.

Table A3: Snack Demand Treatment Effects: Endogenous Prices

	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	-0.028	-0.028	-0.081	-0.077	-0.075	-0.077	-0.070
Median standard error	(0.004)	(0.016)	(0.014)	(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	$\text{WTP}_j > \epsilon_{js}$		$\text{WTP}_j > \epsilon_{js}$		$\text{WTP}_j > \epsilon_{js}$	$\text{WTP}_j > \epsilon_{js}$	$\text{WTP}_j > \epsilon_{js}$
Observed at low price	All	$\text{WTP}_j \leq \epsilon_{js}$		$\text{WTP}_j \leq \epsilon_{js}$		$\text{WTP}_j \leq \epsilon_{js}$	$\text{WTP}_j \leq \epsilon_{js}$	$\text{WTP}_j \leq \epsilon_{js}$

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.

Table A4: Microfinance Application: Estimates of the Effect of Matching and Standard Errors by Correlation Threshold

	-1	-0.2	-0.1	-0.05	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
IV	0.54	1.03	1.08	1.18	1.17	1.24	1.27	1.24	1.22	1.22	1.25	1.23	1.22	1.22
	(0.24)	(0.30)	(0.33)	(0.36)	(0.37)	(0.38)	(0.38)	(0.39)	(0.39)	(0.39)	(0.40)	(0.41)	(0.41)	(0.40)
IV-ARB	1.33	1.3	1.33	1.33	1.35	1.37	1.36	1.34	1.34	1.32	1.32	1.33	1.33	1.33
	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)
	0.1	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.21	0.22	0.23
IV	1.23	1.2	1.3	1.35	1.44	1.52	1.61	1.57	1.58	1.61	1.56	1.54	1.47	1.54
	(0.42)	(0.40)	(0.43)	(0.43)	(0.46)	(0.44)	(0.48)	(0.46)	(0.46)	(0.46)	(0.45)	(0.46)	(0.46)	(0.46)
IV-ARB	(1.34)	(1.32)	(1.32)	(1.31)	(1.31)	(1.31)	(1.30)	(1.27)	(1.29)	(1.33)	(1.34)	(1.32)	(1.30)	(1.27)
	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)	(0.23)

The estimates correspond to the points shown in Figure A9 of the main text. Standard errors are conditional on the selected threshold. All numbers shown are medians over 11 random splits of responses to create the instrument for IV regression.

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 2006; 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);[37] 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).[38]

Stepping away from SP data, portions of the neuroeconomics literature seek to predict choices from neural and/or physiological responses. Smith et al. (2014) focus specifically on passive non-choice neural reactions, and provide proof-of-concept that those types of reactions predict choices.[39] 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; Benjamin et al. 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., 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, Steckel, and Gupta 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. 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,

[37] 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, Bunch, and Train (2000) and Small, Winston, and Yan (2005).

[38] 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).

[39] See also Tusche, Bode, and Haynes (2010) and Levy et al. (2011).

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 (2011a) find that questions concerning likely electoral outcomes (i.e., how others will vote) are better predictors than stated intentions.[40] 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 econometric results

C.1 Proof of Proposition 1

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

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

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

First, we show that the moment condition is valid. By Assumptions 1, and 2,

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

where $\tau^* = \mathbb{E}(Y_j(1) - Y_j(0))$ and $\boldsymbol{\delta}^* \equiv [\boldsymbol{\beta}^T,\ \boldsymbol{\gamma}^T]^T$ with $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ defined in Assumption 2. Similarly, combining the previous assumptions with an unconfoundedness assumption on treatment assignment (Assumption 3A, 3B, or 3D), by the law of iterated expectations:

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

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

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

Second, we derive the expression for the asymptotic variance given in the proposition. Let $\boldsymbol{\psi} = (g^T, \boldsymbol{m}^T)^T$ be the vector stacking the moments.

Define

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

and

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

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

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

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

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

where

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

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

and their derivatives, obtain

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

as given in the proposition.

C.2 Semiparametric identification

While our main estimators make assumptions about functional form, such assumptions are not necessary to identify treatment effects. 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 4. Evaluations overlap. *For each value of the predictors, pooling treatment states, the probability of treatment is bounded away from 0 and 1. Specifically, if $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are the supports of the distributions of predictors in the control and treatment states, respectively, then for all $\boldsymbol{z} \in (\mathbb{Z}_0 \cup \mathbb{Z}_1)$, we have for some $\eta > 0$ at least one of*

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

or

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

Evaluations overlap states that, for any value of the predictors $\boldsymbol{z} \in (\mathbb{Z}_0 \cup \mathbb{Z}_1)$, we observe (a growing number of) settings j for which the hypothetical evaluations corresponding to the realized treatment state coincide with $\boldsymbol{z}$, i.e., $\boldsymbol{Z}_j(W_j) = \boldsymbol{z}$, which may all be in the same treatment state. The overlap assumption is therefore substantially weaker than for standard treatment effects estimators, where for any value of $\boldsymbol{z}$ both the treatment and the control assignment must be possible.. In particular, Assumption 4 can hold even when there is no variation in treatment assignment (Assumption 3A).

With this assumption we can ensure identification:

Proposition 4. *The average effect of the treatment, $\tau = \mathbb{E}(Y_j(1) - Y_j(0))$, is semiparametrically identified under Assumptions 1; one of 3A, 3B, or 3D; and 4.*

Proof. Denote $\boldsymbol{Z}_j(w) = [\boldsymbol{H}_j(w), \boldsymbol{X}_j]$ with $\mathbb{Z}_w$ the support of $\boldsymbol{Z}_j(w)$, for $w \in \{0,1\}$. By the law of iterated expectations, $\mathbb{E}(Y_j(1)) = \mathbb{E}(\mathbb{E}(Y_j(1) \mid \boldsymbol{Z}_j(1)))$. By the evaluations overlap assumption, for any $\boldsymbol{z} \in \mathbb{Z}_1$, either $\Pr(W_j = 1 \mid \boldsymbol{Z}_j(1) = \boldsymbol{z}) > 0$ or $\Pr(W_j = 0 \mid \boldsymbol{Z}_j(0) = \boldsymbol{z}) > 0$ (or both). In the first case, $\mathbb{E}(Y_j(1) \mid \boldsymbol{Z}_j(1) = \boldsymbol{z}) = \mathbb{E}(Y_j(1) \mid \boldsymbol{Z}_j(1) = \boldsymbol{z}, W_j = 1) = \mathbb{E}(Y_j \mid \boldsymbol{Z}_j(1) = \boldsymbol{z}, W_j = 1)$ is identified, where the first equality follows from any version of the unconfoundedness assumption. In the second case, $\mathbb{E}(Y_j(1) \mid \boldsymbol{Z}_j(1) = \boldsymbol{z}) = \mathbb{E}(Y_j(0) \mid \boldsymbol{Z}_j(0) = \boldsymbol{z}) = \mathbb{E}(Y_j(0) \mid \boldsymbol{Z}_j(0) = \boldsymbol{z}, W_j = 0) = \mathbb{E}(Y_j \mid \boldsymbol{Z}_j(0) = \boldsymbol{z}, W_j = 0)$ is identified, where the first equality follows from mapping invariance and the second equality from any version of the unconfoundedness assumption. Hence, $\mathbb{E}(Y_j(1))$ is identified. The argument for $\mathbb{E}(Y_j(0))$ is similar. Hence, τ is semiparametrically identified. □

Proposition 4 says that we can estimate treatment effects without making functional form assumptions. We therefore view parametric assumptions, such as linearity, primarily as useful approximations: our approach is not fundamentally tied to them.

C.3 A LASSO-type estimator for high-dimensional evaluations and non-linear relationships

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

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

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

C.3.1 Description of the estimator

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

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

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

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

Step 1b. Compute approximate balancing weights

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

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, Imbens, and Wager (2018) set the tuning parameter $\zeta = 0.5$ as a default.

Step 2. Estimate the average treatment effect as

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

If we included only the first term in Step 2, the procedure would be analogous to replacing OLS with LASSO in our low-dimensional procedure. The second term in Step 2 addresses the biases associated with high-dimensional estimation and regularization by adding weighted prediction errors from Step 1a. The particular weights $\boldsymbol{\rho}^t$ and $\boldsymbol{\rho}^c$, computed in Step 1b, are meant to reduce estimation errors for $\mathbb{E}\big(\mathbb{E}(Y_j(1) \mid \boldsymbol{Z}_j(1))\big)$ and $\mathbb{E}\big(\mathbb{E}(Y_j(0) \mid \boldsymbol{Z}_j(0))\big)$ in the first term of Step 2, under the assumption of linearity.[41]

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

C.3.2 Theoretical result

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

Proposition 5. *Suppose our Assumptions 1, 2 (here linearity in high-dimensional covariates $\boldsymbol{Z}_j(w)$ rather than $\boldsymbol{H}_j(w)$ and $\boldsymbol{X}_j$); one of 3A, 3B, 3D; and 4, as well as assumptions from Athey, Imbens, and Wager (2018) – exact sparsity Assumption 4, regularity conditions on the covariates $\boldsymbol{Z}$ of Assumption 7, regularity conditions on the (potentially heteroskedastic) regression noise in Corollary 2 – hold. Suppose further that we use the estimator $\hat{\tau}_{\text{arb}}$ with a hard constraint replacing the Lagrange form penalty on the imbalance in our Step 1b (analogous to the constraint in Theorem 2 of Athey, Imbens, and Wager (2018)). Then the estimator $\hat{\tau}_{\text{arb}}$ is asymptotically normal with*

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

where $\hat{V}_{\text{arb}} = \sum_{j=1}^{N} (\rho_j^t - \rho_j^c)^2 (Y_j - \boldsymbol{Z}_j \hat{\boldsymbol{\delta}}_{\text{lasso}})^2$.[42]

Proof. The result follows from Corollary 2 of Athey, Imbens, and Wager (2018) by noting that our unconfoundedness Assumptions 3D, 3A, and 3B each have the same implication as their Assumption 1 for this estimation step, our Assumptions 1, 2, and either of 3D, 3A, or 3B, jointly imply their Assumption 2, and our overlap Assumption 4 is identical to their Assumption 6 after rewriting our variables according to their setup. Their condition on the limit of the odds ratio is not needed in our setting because we observe covariates $\boldsymbol{Z}_j(0)$ and $\boldsymbol{Z}_j(1)$ and an outcome Y_j for all decision problems irrespective of treatment assignment. The two weights $\boldsymbol{\rho}^t$ and $\boldsymbol{\rho}^c$ 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 $\boldsymbol{\rho}^t - \boldsymbol{\rho}^c$ takes the role of their γ in the "Proof of Corollary 6" in their on-line appendix. □

covariates.

[42] In contrast to the variance in Proposition 1, the variance estimator $\hat{V}_{\text{arb}}$ in Proposition 5 is conditional on hypothetical evaluations. Specifically, for a fixed sample size, the weights $(\rho_j^t - \rho_j^c)$ are deterministic (fixed) under sampling of outcomes Y_j conditional on covariates $\boldsymbol{Z}_j$ and treatment assignment W_j. Hence, if one is specifically interested in comparing the estimated standard errors across our low-dimensional and high-dimensional methods, the proper counterpart to $\hat{V}_{\text{arb}}$ from Proposition 5 is the second term of $\hat{V}_{\text{p}}$ from Proposition 1.

C.3.3 Restricting to respondents who are most skilled

When we use ARB in the setup that identifies the most informative respondents in Appendix C.8, we augment the procedure as follows. In Step 1a, rather than using a single penalized (LASSO) regression with regressors corresponding to all thresholds, we estimate a separate regression for each threshold. Because the number of survey responses can be low when restricting to very skilled respondents, we correct for measurement error to address possibly large sampling variation in the remaining responses. The approach we take is similar to splitting the responses in half (within setting) and using the average of the first half as an instrument for the average of the second half, but avoids arbitrary sample splitting. Note that the objective function contains terms $\left(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}\right)^2 = Y_j^2 - 2Y_j\boldsymbol{Z}_j\boldsymbol{\delta} + \boldsymbol{\delta}^T\boldsymbol{Z}_j^T\boldsymbol{Z}_j\boldsymbol{\delta}$. Classical measurement error in $\boldsymbol{Z}_j$ is averaged out across observations for the second term, but squared measurement error in $\boldsymbol{Z}_j^T\boldsymbol{Z}_j$ causes bias. Suppose we have K independent responses from sufficiently skilled respondents for setting j under treatment state W_j, say $\boldsymbol{Z}_{j,k}(W_j)$. Rather than using $(\frac{1}{K}\sum_{k=1}^K \boldsymbol{Z}_{j,k}(W_j))^T(\frac{1}{K}\sum_{k=1}^K \boldsymbol{Z}_{j,k}(W_j))$ to estimate the correctly measured $\boldsymbol{Z}_j^T\boldsymbol{Z}_j$, we use the average of all possible $k \neq k'$ terms $\boldsymbol{Z}_{j,k}(W_j)^T\boldsymbol{Z}_{j,k'}(W_j)$, which yield an unbiased estimate. Averaging over settings j yields an objective function that converges to the population objective function with correctly measured regressors. We estimate separate such "regressions" for each threshold by minimizing the consistent estimate of the residual sum of squares using only respondents who pass the threshold. For a given threshold r, this creates an estimated coefficient vector, say $\hat{\boldsymbol{\delta}}^r$. When selecting one of the thresholds r, Step 1a resembles "subset selection" as an alternative to the LASSO regression in the original version of ARB.

For the purpose of Step 1b, we take the vectors $\boldsymbol{Z}_j(w)$ and $\boldsymbol{Z}_j$ to be the collection of average covariates, concatenating *all* thresholds. That is, Step 1b determines weights that balance the average responses for each threshold. The estimator in Step 2 given the choice of a threshold r^*, is then $\hat{\tau}^{r^*}_{\text{arb}} = \left(\overline{\boldsymbol{Z}^{r^*}(1)} - \overline{\boldsymbol{Z}^{r^*}(0)}\right)\hat{\boldsymbol{\delta}}^{r^*} + \sum_{j=1}^J(\rho_j^t - \rho_j^c)\left(Y_j - \boldsymbol{Z}_j^{r^*}\hat{\boldsymbol{\delta}}^{r^*}\right)$ with $\boldsymbol{Z}_j^{r^*}(w)$ the average evaluations of the respondents passing threshold r^*, and $\overline{\boldsymbol{Z}^{r^*}(w)}$ the average of this variable across settings. The second term in $\hat{\tau}^{r^*}_{\text{arb}}$ensures that the estimate is close to the true effect even if the threshold r^* is not selected correctly in finite samples, as long as the true model is linear in the average hypothetical evaluations under the different thresholds.

Next, we describe a way to choose a threshold. Suppose that, in an infinite sample, we can estimate $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ from Assumption 2 by finding the threshold r^* that minimizes

mean squared error:

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

where $\boldsymbol{Z}_j^r$ are the average evaluations for setting j based on an infinite number of respondents passing threshold r, as well as the intercept and any fixed characteristics. We estimate the squared error of using threshold r in finite samples analogous to the estimation $\boldsymbol{\delta}^r$ above. Specifically, noting that the sample criterion function for r^* includes the average of squared hypothetical evaluations, we estimate the squared evaluations for setting j as the average of $\boldsymbol{Z}_{j,k}^T \boldsymbol{Z}_{j,k'}$ for any two distinct respondents who evaluated setting j under the observed assignment W_j (and are sufficiently skilled given the threshold r). In the simplest case, with just two respondents passing the threshold for each setting, we hence estimate mean squared error as

$$\frac{1}{J}\sum_{j=1}^{J}(Y_j^2 - 2Y_j \frac{\boldsymbol{Z}_{j,1} + \boldsymbol{Z}_{j,2}}{2}\hat{\boldsymbol{\delta}}^r + (\hat{\boldsymbol{\delta}}^r)^T \frac{\boldsymbol{Z}_{j,1}^T \boldsymbol{Z}_{j,2} + \boldsymbol{Z}_{j,2}^T \boldsymbol{Z}_{j,1}}{2}\hat{\boldsymbol{\delta}}^r).$$

C.4 Doubly robust estimators

For an alternative doubly robust estimator along the lines of Chernozhukov et al. (2018) using our Assumptions 1, and either 3D or 3D, the following moment condition satisfies the Neyman orthogonality condition:

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

where $\mu(\boldsymbol{h}, \boldsymbol{x}) = \mathbb{E}(Y_j(0) \mid \boldsymbol{H}_j(0) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) = \mathbb{E}(Y_j(1) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$ is the relationship between outcome and hypothetical evaluations of the realized treatment state, and $e_w(\boldsymbol{h}, \boldsymbol{x}) = \Pr(W_j = w \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$ for $w \in \{0, 1\}$ is the probability that decision problem j is observed in state w conditional on the hypothetical evaluations of that state and fixed characteristics. To avoid biases, μ and e_w should be estimated using cross-fitting. Under suitable conditions for the machine learning estimators of choice for μ and e_w, such a doubly robust estimator may perform well. Note, however, that our framework does not suggest that we are well-positioned to correctly specify a propensity score conditional on hypothetical evaluations. Although this doubly robust moment uses the same structural Assumption 1 to estimate μ, it also requires a standard overlap assumption,

different from the evaluations overlap assumption required for the LASSO-type estimator, bounding conditional treatment probabilities away from 0 and 1 ($e_1(\boldsymbol{h}_1, \boldsymbol{x}) > 0$ for all $(\boldsymbol{h}_1, \boldsymbol{x})$ in the support of $(\boldsymbol{H}(1), \boldsymbol{X})$ and $e_0(\boldsymbol{h}_0, \boldsymbol{x}) > 0$ for all $(\boldsymbol{h}_0, \boldsymbol{x})$ in the support of $(\boldsymbol{H}(0), \boldsymbol{X})$). Consequently, it cannot be used to estimate the effect of a treatment that has not been implemented. It is an interesting question whether it is possible to construct a doubly robust estimator of this type that retains the advantages of our parametric and LASSO-type (residual balancing) estimators.

C.5 Proof of Proposition 2

For ease of notation, include fixed characteristics $\boldsymbol{X}_j$ (except for the intercept) in the hypothetical evaluations $\boldsymbol{H}_j$ and $\boldsymbol{H}_j(w)$ for $w \in \{0, 1\}$. Let $\hat{\boldsymbol{\beta}}$ be the first step regression slope coefficients on $\boldsymbol{H}_j$. We show the result under Assumption 3C with $\tilde{w} = 0$; the proof for $\tilde{w} = 1$ is analogous.

First, for fixed k, note that $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \operatorname{var}(\boldsymbol{H}_j)^{-1} \operatorname{cov}(\boldsymbol{H}_j^T, Y_j)$. Because $Y_j = \alpha + \boldsymbol{H}_j\boldsymbol{\beta} + \epsilon_j$ where $\epsilon_j = \epsilon_j(W_j)$ and α is the intercept, $\operatorname{cov}(\boldsymbol{H}_j^T, Y_j) = \operatorname{var}(\boldsymbol{H}_j)\boldsymbol{\beta} + \operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j)$. Hence, $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \operatorname{var}(\boldsymbol{H}_j)^{-1} \operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j)$. Consider an arbitrary element of the vector $\operatorname{var}(\boldsymbol{H}_j)^{-1} \operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j)$, obtained from the inner product of the corresponding row of $\operatorname{var}(\boldsymbol{H}_j)^{-1}$ with the vector $\operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j)$. For vectors a and b, recall that $\|a^T b\| = \|a\|\|b\| \cos\theta$ where θ is the angle between a and b. Since $\cos\theta \leq 1$, $0 \leq \|a^T b\| \leq \|a\|\|b\|$. Hence, consider $\|\operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j)\|$. From the definitions of $\boldsymbol{H}_j$ and ϵ_j, it immediately follows that $\boldsymbol{H}_j = \boldsymbol{H}_j(0) + W_j(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))$ and $\epsilon_j = \epsilon_j(0) + W_j(\epsilon_j(1) - \epsilon_j(0))$. By linearity, $\operatorname{cov}(\boldsymbol{H}_j^T(0), \epsilon_j(0)) = 0$. Hence, using the expression for $\epsilon_j(1) - \epsilon_j(0)$ shown above:

$$\begin{aligned}\operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j) = &\operatorname{cov}(\boldsymbol{H}_j^T(0), W_j(\epsilon_j(1) - \epsilon_j(0))) \\ &+ \operatorname{cov}(W_j(\boldsymbol{H}_j^T(1) - \boldsymbol{H}_j^T(0)), \epsilon_j(0)) \\ &+ \operatorname{cov}(W_j(\boldsymbol{H}_j^T(1) - \boldsymbol{H}_j^T(0)), W_j(\epsilon_j(1) - \epsilon_j(0)))\end{aligned}$$

By the triangle inequality, $\|\operatorname{cov}(\boldsymbol{H}_j^T, \epsilon_j)\|$ is less than or equal to the sum of the norms of the three terms. By Cauchy-Schwarz, for any element $H_j(0)$ of $\boldsymbol{H}_j(0)$, $\operatorname{cov}(H_j(0), W_j(\epsilon_j(1) -$

$\epsilon_j(0)))^2 \leq \text{var}(H_j(0))\,\text{var}(W_j(\epsilon_j(1) - \epsilon_j(0)))$ and similarly for the other two terms. So

$$|\text{cov}(H_j(0), W_j(\epsilon_j(1) - \epsilon_j(0)))| \leq \text{var}(H_j(0))\sqrt{\frac{\text{var}(W_j(\epsilon_j(1) - \epsilon_j(0)))}{\text{var}(H_j(0))}}$$

$$|\text{cov}(W_j(H_j(1) - H_j(0)), \epsilon_j(0))| \leq \text{var}(H_j(0))\sqrt{\frac{\text{var}(W_j(H_j(1) - H_j(0)))}{\text{var}(H_j(0))}} \cdot \sqrt{\frac{\text{var}(\epsilon_j(0))}{\text{var}(H_j(0))}}$$

$$|\text{cov}(W_j(H_j(1) - H_j(0)), W_j(\epsilon_j(1) - \epsilon_j(0)))| \leq \text{var}(H_j(0))\sqrt{\frac{\text{var}(W_j(H_j(1) - H_j(0)))}{\text{var}(H_j(0))}} \cdot \sqrt{\frac{\text{var}(W_j(\epsilon_j(1) - \epsilon_j(0)))}{\text{var}(H_j(0))}}$$

Next, note that $\text{var}(\boldsymbol{H}_j)^{-1} = \frac{1}{\text{var}(H_j)}$ in the univariate case. Write

$$\text{var}(H_j) = \text{var}(H_j(0)) + \text{var}(W_j(H_j(1) - H_j(0))) + 2\,\text{cov}(H_j(0), W_j(H_j(1) - H_j(0))).$$

So, using Cauchy-Schwarz to bound the covariance, $\text{var}(H_j) \geq \text{var}(H_j(0)) + \text{var}(W_j(H_j(1) - H_j(0)))) - 2\sqrt{\text{var}(H_j(0))}\sqrt{\text{var}(W_j(H_j(1) - H_j(0)))}$. Factor out $\text{var}(H_j(0))$ to get

$$\text{var}(H_j) \geq \text{var}(H_j(0))\Big(1 + \frac{\text{var}(W_j(H_j(1) - H_j(0)))}{\text{var}(H_j(0))} - 2\sqrt{\frac{\text{var}(W_j(H_j(1) - H_j(0)))}{\text{var}(H_j(0))}}\Big).$$

Hence we can bound the norm of $\frac{\text{cov}(H_j, \epsilon_j)}{\text{var}(H_j)}$ using the norm with the right-hand-side above in place of $\text{var}(H_j)$ in the denominator. Then $\text{var}(H_j(0))$ cancels. Finally, under Assumption 3C, the ratios $\frac{\text{var}(W_j(H_j(1) - H_j(0)))}{\text{var}(H_j(0))} \to 0$, $\frac{\text{var}(W_j(\epsilon_j(1) - \epsilon_j(0)))}{\text{var}(H_j(0))} \to 0$, and $\sqrt{\frac{\text{var}(\epsilon_j(0))}{\text{var}(H_j(0))}} < \infty$, so each remaining component of the numerator (after canceling $\text{var}(H_j(0))$) vanishes. In contrast, the remaining denominator, $1 + \frac{\text{var}(W_j(H_j(1) - H_j(0))))}{\text{var}(H_j(0))} - 2\sqrt{\frac{\text{var}(W_j(H_j(1) - H_j(0))))}{\text{var}(H_j(0))}} \to 1$. Hence, the ratio vanishes such that $\lim_{k\to\infty} \text{plim}_{n\to\infty} \hat{\beta}_k - \beta_k = 0$. In the multivariate case, one could add assumptions involving products of the univariate variances and elements of $\text{var}(\boldsymbol{H}_j)^{-1}$ to similarly "cancel" $\text{var}(H_j(0))$, but bounding $\text{var}(\boldsymbol{H}_j)^{-1}$ purely in terms of ratios as in Assumption 3C and marginal variances $\text{var}(H_j(0))$ may not be possible.

For the treatment effect, $\text{plim}_{n\to\infty} \frac{1}{J}\sum_{j=1}^{J} H_j(1) - H_j(0) = \mathbb{E}(H_j(1) - H_j(0))$, so the result of the proposition follows by the continuous mapping theorem. For proportional error, given $\tau_k \neq 0$, $\text{plim}_{n\to\infty} \frac{\hat{\tau}_k}{\tau_k} = \frac{\text{plim}_{n\to\infty} \hat{\tau}_k}{\mathbb{E}(H_j(1) - H_j(0))\beta}$ and $\text{plim}_{n\to\infty} \hat{\tau}_k = \mathbb{E}(H_j(1) - H_j(0))\,\text{plim}_{n\to\infty} \hat{\beta}_k$ (by the

continuous mapping theorem), so $\lim_{k\to\infty} \operatorname{plim}_{n\to\infty} \hat{\beta}_k = \beta$ yields $\lim_{k\to\infty} \operatorname{plim}_{n\to\infty} \frac{\hat{\tau}_k}{\tau_k} = 1$ implying the result.

C.6 Partial identification

C.6.1 Semiparametric partial identification

We state a semiparametric partial identification result here. Inspection of the proof reveals that the bounds are sharp: For any point between the bounds, including the end points, there exists a joint distribution of $(Y(0), Y(1), W, \boldsymbol{H}(0), \boldsymbol{H}(1), \boldsymbol{X})$ that is consistent with the distribution of the observable $(Y, W, \boldsymbol{H}(0), \boldsymbol{H}(1), \boldsymbol{X})$ and yields that point as the average treatment effect. In the main text, we give bounds, additionally assuming linearity, as the solution to a linear programming formulation that is fast to compute even with continuous evaluations and fixed characteristics.

Proposition 6. *Suppose Assumption 1 holds and potential outcomes are bounded,* $\underline{Y} \leq Y_j(w) \leq \bar{Y}$ *with* $\underline{Y}, \bar{Y} \in \mathbb{R}$ *for* $w = 0, 1$ *and all* j*. Then the average treatment effect is bounded by* $\tau_{lb} \leq \tau \leq \tau_{ub}$ *with bounds defined as*

$$\tau_{lb} = \int\int \tilde{\mu}_{lb}(\boldsymbol{h}, \boldsymbol{x})(f_{(\boldsymbol{H}(1),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) - f_{(\boldsymbol{H}(0),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}))\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$$
$$\tau_{ub} = \int\int \tilde{\mu}_{ub}(\boldsymbol{h}, \boldsymbol{x})(f_{(\boldsymbol{H}(1),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) - f_{(\boldsymbol{H}(0),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}))\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$$

where

$$\tilde{\mu}_{lb}(\boldsymbol{h}, \boldsymbol{x}) = \begin{cases} \mu_{lb}(\boldsymbol{h}, \boldsymbol{x}) & \text{if } f_{(\boldsymbol{H}(1),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) > f_{(\boldsymbol{H}(0),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) \\ \mu_{ub}(\boldsymbol{h}, \boldsymbol{x}) & \text{otherwise} \end{cases}$$
$$\tilde{\mu}_{ub}(\boldsymbol{h}, \boldsymbol{x}) = \begin{cases} \mu_{ub}(\boldsymbol{h}, \boldsymbol{x}) & \text{if } f_{(\boldsymbol{H}(1),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) > f_{(\boldsymbol{H}(0),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) \\ \mu_{lb}(\boldsymbol{h}, \boldsymbol{x}) & \text{otherwise} \end{cases}$$
$$\mu_{lb}(\boldsymbol{h}, \boldsymbol{x}) = \max_{w\in\{0,1\}} \Big\{ \Pr(W_j = w \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) \mathbb{E}(Y_j \mid W_j = w, \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) + (1 - \Pr(W_j = w \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}))\underline{Y} \Big\}$$
$$\mu_{ub}(\boldsymbol{h}, \boldsymbol{x}) = \min_{w\in\{0,1\}} \Big\{ \Pr(W_j = w \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) \mathbb{E}(Y_j \mid W_j = w, \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) + (1 - \Pr(W_j = w \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}))\bar{Y} \Big\}$$

Proof. By the law of iterated expectations, the average treatment effect equals $\tau =$

$\mathbb{E}(\mathbb{E}(Y_j(1) \mid \boldsymbol{H}_j(1), \boldsymbol{X}_j)) - \mathbb{E}(\mathbb{E}(Y_j(0) \mid \boldsymbol{H}_j(0), \boldsymbol{X}_j))$. Writing the outer expectations in integral form and using mapping invariance, we can write the treatment effect as $\tau = \int\int \mu(\boldsymbol{h}, \boldsymbol{x})(f_{(\boldsymbol{H}(1).\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) - f_{(\boldsymbol{H}(0).\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}))\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$, where the densities of $(\boldsymbol{H}(1), \boldsymbol{X})$ and $(\boldsymbol{H}(0), \boldsymbol{X})$, $f_{(\boldsymbol{H}(1),\boldsymbol{X})}$ and $f_{(\boldsymbol{H}(0),\boldsymbol{X})}$, are identified from the data, and $\mu(\boldsymbol{h}, \boldsymbol{x}) = \mathbb{E}(Y_j(0) \mid \boldsymbol{H}_j(0) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) = \mathbb{E}(Y_j(1) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$. We can bound either conditional expectation by combining the identified $\mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}, W_j = w)$ with the bound $\underline{Y} \leq \mathbb{E}(Y_j(w) \mid \boldsymbol{H}_j(w) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}, W_j = 1 - w) \leq \bar{Y}$. The identifying power of the mapping invariance assumption is that we need only consider the *intersection* of the resulting bounds on $\mathbb{E}(Y_j(1) \mid \boldsymbol{H}_j(1) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$ and $\mathbb{E}(Y_j(0) \mid \boldsymbol{H}_j(0) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x})$ as a bound for $\mu(\boldsymbol{h}, \boldsymbol{x})$, which can be noticeably tighter than the separate bounds.[43] From the integral equation above, the upper bound for τ then uses the upper bound on μ whenever $f_{(\boldsymbol{H}(1).\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) > f_{(\boldsymbol{H}(0).\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x})$ and the lower bound on μ whenever $f_{(\boldsymbol{H}(1).\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) < f_{(\boldsymbol{H}(0).\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x})$. The lower bound for τ does the reverse. In the main text, we operationalize this analysis for continuous evaluations and fixed characteristics by also imposing linearity. □

C.6.2 Discretization

Under mapping invariance and linearity, in the population $\mathbb{E}(\boldsymbol{Z}_j(0)^T\boldsymbol{Z}_j(0))^{-1}\mathbb{E}(\boldsymbol{Z}_j(0)^T Y_j(0)) = \mathbb{E}(\boldsymbol{Z}_j(1)\boldsymbol{Z}_j(1))^{-1}\mathbb{E}(\boldsymbol{Z}_j(1)^T Y_j(1))$ where $\boldsymbol{Z_j}(w)$ contains hypothetical evaluations, fixed characteristics, and the intercept. Note that the infeasible OLS estimator, $(\boldsymbol{Z}(0)^T\boldsymbol{Z}(0))^{-1}\boldsymbol{Z}(0)^T\boldsymbol{Y}(0)$, is numerically equivalent to $(\boldsymbol{Z}(0)^T\boldsymbol{Z}(0))^{-1}\boldsymbol{Z}(0)^T\tilde{\boldsymbol{Y}}(0)$ where $\tilde{Y}_j(0)$ equals the observed outcome of observation j if it is not treated, and equals the average of (unobserved) untreated potential outcomes of observations with the same regressors $\boldsymbol{Z}(0)$ otherwise:

$$\tilde{Y}_j(0) = \begin{cases} Y_j(0) & \text{if } W_j = 0 \\ \frac{\sum_{j'=1}^{J} W_{j'} 1\{\boldsymbol{Z}_j(0) = \boldsymbol{Z}_{j'}(0)\} Y_{j'}(0)}{\sum_{j'=1}^{J} W_{j'} 1\{\boldsymbol{Z}_j(0) = \boldsymbol{Z}_{j'}(0)\}} & \text{if } W_j = 1 \end{cases}$$

and similarly for treated potential outcomes. The average used for $\tilde{Y}_j(0)$ in the case where $W_j = 1$ is unobserved and can range between $\underline{Y}$ and $\bar{Y}$. If the regressors are discrete, the number of distinct averages of this kind is at most the number of support points of the regressor and hence does not grow with the sample size. If the regressors are continuous, we form discrete groups based on the covariates and, if $W_j = 1$, take $\tilde{Y}_j(0)$ as fixed within group.

[43] If the intersection is empty at $(\boldsymbol{h}, \boldsymbol{x})$, mapping invariance is rejected at this point. If $\Pr(W_j = 1 \mid \boldsymbol{H}_j(1) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) = 1$ or $\Pr(W_j = 1 \mid \boldsymbol{H}_j(0) = \boldsymbol{h}, \boldsymbol{X}_j = \boldsymbol{x}) = 0$, then $\mu(\boldsymbol{h}, \boldsymbol{x})$ is point identified (unless mapping invariance is rejected).

For a sufficiently large number of groups, the approximation tends to have little effect. Hence, inference results that allow for a growing number of parameters and constraints may offer a reasonable approximation to the continuous regressor case when the number of groups grows (slowly) with the sample size.

In practice, we discretize the covariate space using a "regression tree" based on greedy splits. Covariate values that fall into the same leaf are in the same group. A benefit of this tree-based discretization is that it is non-stochastic, whereas, for instance, the k-means algorithm typically depends on a random initial allocation. The nodes of the tree partition to covariate space based on the value of a covariate being above or below a threshold. At each node, we consider each covariate and each possible threshold. We choose the split (covariate and threshold) that minimizes the total squared error. For a given split that we consider, within each resulting leaf and across observations and covariates, we calculate the sum of squared deviations from the within-leaf mean, and sum the result across leaves. We avoid splits that would yield a leaf containing only a single observation. In practice, we face a trade-off between choosing a large number of groups such that the discretized problem better approximates the original problem, and choosing a small number of groups such that inference procedures based on the number of groups growing only slowly with the sample size offer a reasonable approximation.

Continuing the bounding exercise in Section 6.2 in footnote 36: when discretizing the 300 observations into 60 groups, imposing $\underline{Y} = 0$, $\bar{Y} = 14.87$, mapping invariance, and linearity (but not $\underline{\tau} = 0$), the bounds to the discretized problem ($[0.47, 1.78]$) are roughly similar to the original problem, and a 90% confidence interval using the approach of Fang et al. (2023) is given by $[0.31, 2.72]$; a 95% confidence interval is $[0.20, 2.84]$. Adding $\underline{\tau} = 0$, even with 120 groups the discretized problem (bounds $[1.12, 1.19]$) does not resemble the original problem.

C.7 Sample selection model

C.7.1 Description

To understand the intuition for this procedure, imagine taking only the subset of settings assigned one of the two treatment states, and regressing outcomes on hypotheticals. Correlations between the hypotheticals and the likelihood of observing a setting in that treatment state potentially produce bias that is treatable through a standard sample selection correction.

Assumption (3E). **Sample selection model.** *For* $w \in \{0, 1\}$,

$$Y_j(w) = \boldsymbol{H}_j(w)\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \epsilon_j(w)$$

with $\mathbb{E}(\epsilon_j(w) \mid \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j) = 0$*, and the treatment is selected according to*

$$W_j = 1\{\boldsymbol{H}_j(0)\boldsymbol{\alpha}_0 + \boldsymbol{H}_j(1)\boldsymbol{\alpha}_1 + \boldsymbol{X}_j\boldsymbol{\alpha}_x + \eta_j > 0\}$$

where (i) $\epsilon_j(0), \epsilon_j(1), \eta_j \perp\!\!\!\perp \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j$*, (ii)* $\mathbb{E}(\epsilon_j(w)) = 0$*, (iii)* $\boldsymbol{\alpha}_0 \neq \mathbf{0}$*,* $\boldsymbol{\alpha}_1 \neq \mathbf{0}$*,* $\boldsymbol{\alpha}_x \neq \mathbf{0}$*, and (iv)* $\epsilon_j(0), \epsilon_j(1), \eta_j$ *are jointly normally distributed with the variance of* η_j *normalized to* 1 *and the covariance between* η_j *and* $\epsilon_j(w)$ *defined as* $\sigma_{\epsilon(w),\eta}$ *for* $w = 0, 1$.

Note that the correlation of $\epsilon_j(w)$ and η_j is unrestricted, allowing treatment to depend on unobservables. While we state the assumption here with normal errors, the usual extensions to cases with nonparametric errors are possible at the cost of more challenging estimation in small samples.

To estimate $\boldsymbol{\beta}$ under Assumption 3E, one follows these steps: First, use a Probit regression of W_j on $\boldsymbol{H}_j(0)$, $\boldsymbol{H}_j(1)$, and $\boldsymbol{X}_j$ to calculate fitted values $\hat{W}_j = \boldsymbol{H}_j(0)\hat{\boldsymbol{\alpha}}_0 + \boldsymbol{H}_j(1)\hat{\boldsymbol{\alpha}}_1 + \boldsymbol{X}_j\hat{\boldsymbol{\alpha}}_x$ of the linear index. Second, generate the new regressors $\hat{V}_j(0) = \frac{(1-W_j)\phi(\hat{W}_j)}{1-\Phi(\hat{W}_j)}$ and $\hat{V}_j(1) = \frac{W_j\phi(\hat{W}_j)}{\Phi(\hat{W}_j)}$. Third, regress Y_j on $\boldsymbol{H}_j$, $\boldsymbol{X}_j$, $\hat{V}_j(0)$, and $\hat{V}_j(1)$. Finally, use the estimated coefficient on $\boldsymbol{H}_j$, $\hat{\boldsymbol{\beta}}$, from the third step to calculate $\hat{\tau}_{\text{ssm}} = \frac{1}{J}\sum_{j=1}^{J}(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\hat{\boldsymbol{\beta}}$. Effectively, the hypothetical evaluation of the unobserved treatment state serves as an internal instrument so that identification does not rely exclusively on the functional form. An additional adjustment to the final step is necessary when estimating the ATC or ATT instead of the ATE; see Appendix C.7.3.

Proposition 7. *Suppose the data* $(Y_j, W_j, \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j)_{j=1}^{J}$ *are a random sample of independent observations and standard regularity conditions hold. Under Assumption 3E, the parametric estimator* $\hat{\tau}_{ssm}$ *is consistent for the average treatment effect* τ *and is asymptotically normal:*

$$\sqrt{J}\left(\hat{\tau}_{ssm} - \tau\right) \to \mathcal{N}\left(0, V_{\tau,ssm}\right)$$

with a formula for $V_{\tau,ssm}$ *given at the end of the proof.*

In the microfinance application, the estimate of the ATC is relatively insensitive to the particular assumption on treatment assignment used to obtain point identification. As discussed, the estimates presented in Table 3 Columns (5)–(11) are valid under Unconfoundedness with the correctly specified linear model. For instance, the specification of

Column (8) yields a point estimate of 1.28 (bootstrap s.e. 0.39). Using the same specification of hypotheticals and fixed characteristics but adapting the method to Assumption 3E, we obtain 1.31 (0.46).

C.7.2 Proof of Proposition 7

Define the "single index" $V_j = \boldsymbol{H}_j(0)\boldsymbol{\alpha}_0 + \boldsymbol{H}_j(1)\boldsymbol{\alpha}_1 + \boldsymbol{X}_j\boldsymbol{\alpha}_x$. To find an estimable equation involving the parameter $\boldsymbol{\beta}$, we use that

$$\mathbb{E}(Y_j \mid \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j, W_j = w) = \boldsymbol{H}_j(w)\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \mathbb{E}(\epsilon_j(w) \mid \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j, W_j = w).$$

Conditioning on $W_j = 1$ is equivalent to conditioning on the event $\eta_j \geq -V_j$. Using the properties of the truncated normal distribution, $\mathbb{E}(\epsilon_j(1) \mid \eta_j \geq -V_j) = \sigma_{\epsilon(1),\eta}\frac{\phi(V_j)}{\Phi(V_j)}$, where the ratio is the inverse Mills ratio. Conditioning on $W_j = 0$ is equivalent to conditioning on the event $\eta_j < -V_j$, and, similarly to the $\eta_j \geq -V_j$ case, $\mathbb{E}(\epsilon_j(0) \mid \eta_j < -V_j) = -\sigma_{\epsilon(0),\eta}\frac{\phi(V_j)}{1-\Phi(V_j)}$. Hence,

$$\mathbb{E}(Y_j \mid \boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j, W_j) = \boldsymbol{H}_j(w)\boldsymbol{\beta} + \boldsymbol{X}_j\boldsymbol{\gamma} + \sigma_{\epsilon(1),\eta}\frac{W_j\phi(V_j)}{\Phi(V_j)} - \sigma_{\epsilon(0),\eta}\frac{(1-W_j)\phi(V_j)}{1-\Phi(V_j)}.$$

Define $\hat{W}_j = \boldsymbol{H}_j(0)\hat{\boldsymbol{\alpha}}_0 + \boldsymbol{H}_j(1)\hat{\boldsymbol{\alpha}}_1 + \boldsymbol{X}_j\hat{\boldsymbol{\alpha}}_x$ with $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_x)$ the coefficients from a probit regression of W_j on $\boldsymbol{H}_j(0)$, $\boldsymbol{H}_j(1)$, and $\boldsymbol{X}_j$, and let $\hat{V}_j(1) = \frac{W_j\phi(\hat{W}_j)}{\Phi(\hat{W}_j)}$ and $\hat{V}_j(0) = \frac{(1-W_j)\phi(\hat{W}_j)}{1-\Phi(\hat{W}_j)}$. Then the coefficients on $\boldsymbol{H}_j(W_j)$ in a regression of Y_j on $\boldsymbol{H}_j(W_j)$, $\boldsymbol{X}_j$, $\hat{V}_j(1)$, and $\hat{V}_j(0)$ consistently estimates $\boldsymbol{\beta}$. For the average treatment effect, consistent estimation of $\hat{\boldsymbol{\beta}}$ combined with the consistent sample mean estimator for $\mathbb{E}(H_j(1) - H_j(0))$ yields a consistent estimator of the average effect because $\mathbb{E}(Y_j(1) - Y_j(0)) = \mathbb{E}(H_j(1) - H_j(0))\boldsymbol{\beta}$ as before.

For the variance, write the moment functions as

$$\begin{aligned}
g(y, \boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}, \boldsymbol{r}, \tau, \boldsymbol{\delta}, \boldsymbol{\alpha}) &= \tau - (\boldsymbol{z}_1 - \boldsymbol{z}_0)\boldsymbol{\delta} \\
\boldsymbol{m}(y, \boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}, \boldsymbol{r}, \tau, \boldsymbol{\delta}, \boldsymbol{\alpha}) &= [\boldsymbol{z}, \frac{w\phi(\boldsymbol{r}\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{r}\boldsymbol{\alpha})}, \frac{(1-w)\phi(\boldsymbol{r}\boldsymbol{\alpha})}{\Phi(\boldsymbol{r}\boldsymbol{\alpha})}]^T(y - \boldsymbol{z}\boldsymbol{\delta} - \sigma_1\frac{w\phi(\boldsymbol{r}\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{r}\boldsymbol{\alpha})} - \sigma_0\frac{(1-w)\phi(\boldsymbol{r}\boldsymbol{\alpha})}{\Phi(\boldsymbol{r}\boldsymbol{\alpha})}). \\
\boldsymbol{q}(y, \boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}, \boldsymbol{r}, \tau, \boldsymbol{\delta}, \boldsymbol{\alpha}) &= \boldsymbol{r}^T(w\frac{\phi(\boldsymbol{r}\boldsymbol{\alpha})}{\Phi(\boldsymbol{r}\boldsymbol{\alpha})} - (1-w)\frac{\phi(\boldsymbol{r}\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{r}\boldsymbol{\alpha})})
\end{aligned}$$

The moment function g corresponds to the final treatment effect estimation. The moment functions $\boldsymbol{m}$ correspond to the regression of the outcome on the regressors including terms involving the inverse Mills ratio. The moment functions $\boldsymbol{q}$ correspond to probit

regression to estimate the coefficients of the linear index. Let $\boldsymbol{\psi}^*$ denote the vector of moment functions stacking g, $\boldsymbol{m}$, and $\boldsymbol{q}$, evaluated at $y = Y_j$, $\boldsymbol{z}_w = [\boldsymbol{H}_j(w), \boldsymbol{X}_j]$, $\boldsymbol{z} = W_j\boldsymbol{z}_1 + (1 - W_j)\boldsymbol{z}_0$, $\boldsymbol{r} = [\boldsymbol{H}_j(0), \boldsymbol{H}_j(1), \boldsymbol{X}_j]$ and the true parameter values. Let $\nabla\boldsymbol{\psi}^*$ similarly denote the matrix of derivatives of the moment functions with respect to the parameters, evaluated at the random variables and true parameters as before. Then the asymptotic variance of the GMM estimator $(\hat{\tau}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\alpha}})$ of the vector of true parameter values is given by $V_{\text{ssm}} = \mathbb{E}(\nabla\boldsymbol{\psi}^*)^{-1}\mathbb{E}(\boldsymbol{\psi}^*)\mathbb{E}(\nabla\boldsymbol{\psi}^{*,T})^{-1}$ (Newey and McFadden 1994). The $(1,1)$ element of V_{ssm} is the asymptotic variance of the treatment effect estimator, $V_{\tau,\text{ssm}}$.

C.7.3 Estimating the ATC

To estimate the average effect of the treatment on the control (ATC), $\mathbb{E}(Y_j(1) - Y_j(0) \mid W_j = 0)$, note that $\mathbb{E}(Y_j(0) \mid W_j = 0)$ can be estimated consistently by taking the sample analog. Hence, we focus on estimation of $\mathbb{E}(Y_j(1) \mid W_j = 0) = \mathbb{E}(\boldsymbol{H}_j(1) \mid W_j = 0)\boldsymbol{\beta} + \mathbb{E}(\epsilon(1) \mid W_j = 0)$. For $\mathbb{E}(\boldsymbol{H}_j(1) \mid W_j = 0)$, the sample analog estimator is consistent. For $\mathbb{E}(\epsilon_j(1) \mid W_j = 0)$, however, note that in contrast to the case under unconfoundedness assumption, we cannot use that $\mathbb{E}(\epsilon_j(1) \mid W_j = 0) = \mathbb{E}(\mathbb{E}(\epsilon_j(1) \mid \boldsymbol{H}_j(1), W_j = 0) \mid W_j = 0)$ and then appeal to unconfoundedness to argue that $\mathbb{E}(\epsilon_j(1) \mid \boldsymbol{H}_j(1), W_j = 0) = \mathbb{E}(\epsilon_j(1) \mid \boldsymbol{H}_j(1))$ and finally the linear model assumption to get $\mathbb{E}(\epsilon_j(1) \mid \boldsymbol{H}_j(1)) = 0$. Instead, we use the truncated normal calculations to find

$$\mathbb{E}(\epsilon_j(1) \mid W_j = 0) = \mathbb{E}(\mathbb{E}(\epsilon_j(1) \mid \eta_j < V_j) \mid W_j = 0) = -\sigma_{\epsilon(1),\eta}\mathbb{E}\left(\frac{\phi(V_j)}{\Phi(V_j)} \mid W_j = 0\right).$$

Here, $\sigma_{\epsilon(1),\eta}$ is consistently estimated by the coefficient on $\frac{W_j\phi(\hat{W}_j)}{\Phi(\hat{W}_j)}$ in the first step regression, and we can consistently estimate $\mathbb{E}\left(\frac{\phi(V_j)}{\Phi(V_j)} \mid W_j = 0\right)$ by $\frac{1}{\sum_{j=1}^{J}(1-W_j)}\sum_{j=1}^{J}(1 - W_j)\frac{\phi(\hat{W}_j)}{\Phi(\hat{W}_j)}$.

C.8 Using the hypothetical response from those who are most skilled

Differences between the populations making real and hypothetical evaluations may make the hypothetical evaluations less predictive of the choices. In the 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.[44]

A potential strategy for improving upon our basic estimator is to use only the hypothetical evaluations of respondents who are best able to predict real outcomes. This strategy requires us to elicit a prediction of real outcomes as one of the hypothetical evaluations. We define the *latent response quality*, r_{kj}, for respondent k's evaluations of setting j as the correlation between k's predictions and outcomes *for other settings* $j' \neq j$. For the purpose of estimating treatment effects, one can set a quality threshold r^* and drop all observations with latent quality below this threshold, $r_{kj} \leq r^*$.[45] Because this strategy reduces the number of evaluations per setting, it may be appropriate to remove any bias due to classical measurement error in hypothetical evaluations by replacing Step 1 of our method with an estimator that corrects for measurement error using repeated measurements, similar to instrumental variables, as detailed in Appendix C.3.3. By varying r^*, the analyst can check whether poor quality evaluations drive the results.

We illustrate this strategy by filtering respondents based on correlations between their "quintile projections" and actual fundraising velocities for loan profiles displayed in their actual treatment states.[46] Figure A9 shows (in orange squares) how the estimates vary with the correlation threshold r^* used for filtering responses. For each correlation threshold, we estimate a model of the outcome on the quintile projection and use it to derive treatment effects. These estimates vary substantially based on the correlation threshold, from 0.39 to 1.14, and tend toward the ground truth estimate as we limit the analysis to responses with higher latent quality. However, this estimator raises two issues: First, which correlation threshold should be used? Because estimates are highly correlated across thresholds, it can be difficult to select among them. Second, when the quintile projection only enters linearly, the functional form may be misspecified (Assumption 2 may be violated). Because the distribution of quintile projections differs across treatment states, such misspecification can cause differential prediction error for treated and untreated outcomes in our second step.

A second method may address both of these issues. For each threshold r, this method uses an approach derived from ARB that balances residuals based on data for all thresholds. With respect to the first issue, the selection of a threshold resembles the selection of covariates by the LASSO, for which ARB yields greater robustness against finite sample

[44] 10% state they have made one loan using the website, and a little over 3% state they have made two or more loans.

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

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

mistakes. With respect to the second issue, ARB adds residuals that reflect divergences between the correct functional form and the linear model, much as it potentially corrects LASSO predictions if the LASSO incorrectly drops higher order terms in finite samples. See Appendix C.3 for a detailed description of the algorithm. As the purple dots in Figure A9 show, the resulting estimates are much less sensitive to the choice of the threshold: all of them are between 1.28 and 1.32, only slightly higher than ground truth (1.24).

The issue of mismatch between the populations for real choices and hypothetical responses does not arise in our snack application because we recruited the participants who make hypothetical choices from the same population as the participants who make real choices. Nevertheless, for completeness, we replicate the analysis of this section for the snack application in Appendix Figure A10.

C.9 Relation to other methods

C.9.1 Linear factor models and synthetic controls

Linear factor models can provide an alternative microfoundation for our estimators. Suppose setting j under treatment state w induces the menu of motivational attribute bundles $\boldsymbol{\Theta}_j(w)$. The outcome and hypothetical evaluations are linear functions of these latent "factors":

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

where $\boldsymbol{H}_j(w) = [H_j^1(w), \ldots, H_j^{Q_H}(w)] \in \mathbb{R}^{Q_H}$, such that $H_j^q(w)$ is the aggregate response to the q^{th} hypothetical question.[47] Only Y and $\boldsymbol{H}$ are observed; latent variables include $\boldsymbol{\Theta}_j(w)$ (a row vector) and weights $\boldsymbol{\phi}_Y$ and $\boldsymbol{\phi}_{H^q}$ (column vectors). Factor models are also often used to derive properties of synthetic control methods (Abadie, Diamond, and Hainmueller 2010). To see that the methods are similar, suppose j is the time period and $H_j^q(w)$ is the outcome for the q^{th} "donor unit" in period j. In the synthetic control method, one regresses the control potential outcome of the treated unit in period j, $Y_j(0)$, on contemporaneous control potential outcomes of donor units, $\boldsymbol{H}_j(0)$, using periods where these outcomes are observed (pre-treatment). This is analogous to Step 1 of our method using the observed outcomes $Y_j(W_j)$, except that synthetic controls often restrict the coefficients to be positive and to sum to 1. Synthetic control then uses the relationship to predict counterfactual

[47] For simplicity, we consider a model without fixed characteristics $\boldsymbol{X}_j$; including them is straightforward.

outcomes $\hat{Y}_j(1 - W_j)$ for the (post-treatment) periods where the control potential outcomes are unobserved, and takes the difference $Y_j(1) - \hat{Y}_j(0)$ between observed and predicted counterfactual, "synthetic," outcome in post-treatment periods. Our Step 2 instead uses the relationship to predict both unobserved outcomes *and* observed outcomes, taking the average of the differences $\hat{Y}_j(1) - \hat{Y}_j(0)$ across all observations. The difference between our method and the synthetic control method in Step 2 arises primarily because our assumptions allow us to model both treated and control potential outcomes (whereas the synthetic control method only models control potential outcomes using the factor structure). This difference also allows us to focus on different estimands, including the average treatment effect.

C.9.2 Statistical surrogates

We assume that treatments affect the outcomes of interest only through psychological motivations. Consequently, we treat hypothetical evaluations much like statistical surrogates (Prentice 1989). This literature has recently received renewed interest in economics in the context of estimating the effects of a treatment on long-term outcomes using short-term outcomes as surrogates (Athey, Chetty, and Imbens 2020; Athey et al. forthcoming). However, statistical surrogates are observed only for the realized treatment state, whereas we observe hypothetical evaluations for all treatment states. This distinction leads to different assumptions, estimators, and properties.

D Snack Demand Application

D.1 Groups

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

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

Group M (35 subjects): Subjects considered the same choice problems as in group R, but were told in advance that all but five decisions would be hypothetical. The five real choices were interspersed among the hypothetical choices, but clearly indicated when they were presented. For each subject, the five items were drawn at random from a larger group of fifteen, selected for their representativeness,[48] and each was offered at a price of 75 cents. The purpose of this "mixed" group is to investigate the concern that the low probability with which any given choice problem was implemented in group R influenced purchase frequencies (e.g., if subjects treated the "real" choices as hypothetical).

Group HCT (28 subjects): Subjects performed that same task as in group H, but a "cheap talk" script (as in Cummings and Taylor 1999) was added to the experimental instructions, with the objective of inducing subjects to take the hypothetical choices more seriously, and thereby minimize hypothetical bias.[49]

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

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

Group HWTP (28 subjects): Subjects expressed a hypothetical willingness to pay (WTP) for all of the food items, each of which appeared only once. We employed this protocol because much of the literature explores the accuracy of hypothetical WTPs rather than binary choices. We used the same subjects for groups HWTP and L (below).[50]

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

[49] 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).

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

Group SWB (28 subjects): For each potential outcome, subjects indicated their anticipated subjective well-being: "How happy would you be if you received this item (and ONLY this item) to eat as a snack during the second part of this experiment, and a price of $X was deducted from your show-up payment?" (with 1="very unhappy" and 7="very happy"). Each item appeared twice, once with a price of 25 cents and once with a price of 75 cents.

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

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

Group S (29-38 subjects):[51] 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

[51]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.

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
- 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," which serves as the left-out baseline)
- the average vicarious choice likelihood (on a 5 point scale)
- the fraction of respondents stating a vicarious likelihood of at least each level (except for "very unlikely").

Second, variables that are not price-specific; for each of the six questions of Group S (see Appendix 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 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" group, in which subjects were told that five of their choices would be real (that is, one of the five would be chosen at random and implemented), and the rest would be hypothetical. The real choices were clearly identified and interspersed among the hypothetical ones. In that group, the implementation probability for each real choice was 1 in 5 rather than 1 in 378. We elicited 175 real choices through this "mixed" group,

pertaining to 15 distinct items (at a price of $0.75). We then pooled that data with 450 choices involving the same 15 items from the "real choice" group, and estimated a logistic regression relating the purchase decision to a set of 15 product dummies as well as a "mixed choice group" dummy. If the "real choice group" subjects viewed their choices as real, the coefficient for the "mixed choice group" dummy should be zero; if they viewed those choices as partially hypothetical, then the "mixed choice group" coefficient should be negative given the documented direction of hypothetical bias. In fact, it was positive 0.11, with a standard error of 0.21 (assuming independent observations). The difference is both statistically insignificant and of an economically small magnitude (average marginal effect of less than 2 percentage points). The coefficient indicates that the purchase frequencies were, if anything, slightly higher for real choices in the "mixed choice" group than in the "real choice" group, which is inconsistent with the hypothesis that participants in the "real choice" group were more inclined to view their choices as hypothetical than were participants in the "mixed choice" group.

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 $\sigma_\omega^2 = \text{var}(\omega_j) = \text{var}(\mathbb{E}(\omega_j \mid \mu_j)) + \mathbb{E}(\text{var}(\omega_j \mid \mu_j))$. The conditional expectation in the first term is 0 because $\mathbb{E}(H_j \mid \mu_j) = \mu_j$. 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 $\text{var}(\omega_j \mid \mu_j) = \mu_j(1-\mu_j)/N$. 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. Similarly, Carson, Groves, and List (2011) found that the variance of valuations rises when choices become less consequential.

D.5 Overlap assessment

For all applications, we recommend producing overlap plots such as those shown in Figure A3 to assess the potential stability of equation 1. The figure compares the four hypothetical choice variables and the dichotomized WTP variable. The hypothetical choices are predictors that may be suited to any application, while the WTP choice is mostly applicable for studying price variation, or measuring "contingent valuation." Part (a) of the figure depicts the distributions of evaluations for the high price in purple and for the low price in orange. The analyst can generate such overlap plots in any application, even without observing ground truth.

The upper left panel of part (a) focuses on the standard hypothetical choice variable. The distribution of this variable with the low price overlaps the distribution with the high price (almost) completely, and vice versa. Overlap is also reasonably complete for the hypothetical choices based on the cheap-talk script, own choice likelihood ('intensity'), and vicarious choice likelihood.

In contrast, overlap for the dichotomized WTP choice variable is asymmetric, as shown in the upper right panel. While the distribution of the WTP choice with the high price largely overlaps the distribution at the low price, the opposite is not true. Consider the region of this variable below 0.4: while about half of all snacks fall into this range when priced at \$0.75, there are no snacks in this range when priced at \$0.25. As a result, if we were to observe all real choices at the low price, predicting purchases at the high price based on WTP choice would require extrapolation significantly beyond the range of observation. Because overlap for the WTP variable is limited, we exclude it from our multivariate specifications throughout.

Because our experiment reveals ground truth, we can further diagnose the overlap problem with hypothetical WTP. Part (b) of the figure uses observations of actual demand at both prices to show that the predictive relationship may be approximately linear for one measure (standard hypothetical choice) but not for another (WTP choice, which exhibits nonlinearity at lower values). In practice, if we observed all snacks at the low price, we would only be able to plot the orange squares, from which we might infer the orange curves. Because the low-price data does not span hypothetical WTP purchase frequencies below 0.4, it cannot reveal that the relationship becomes markedly non-linear over that range. We can uncover this property in our experiment (for which we actually have real choices at both prices) by inspecting the high-price data (the purple curve).

D.6 Price setting simulation

For the price setting simulation described in Section 3.3, we define profit as $\pi_j(w) = (w \cdot 0.75 + (1-w) \cdot 0.25 - c)Y_j(w)$ for $w \in \{0,1\}$ for snack j and average profit as $\bar{\pi}(\boldsymbol{w}^*) = \frac{1}{J}\sum_j \left[(w_j^* - c)Y_j(w_j^*)\right]$. We set marginal costs c so that it is optimal to sell half of the snacks at the low price and half at the high price. 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. We assume the producer observes demand for snack j at a single price W_j, predicts demand at the other price, $\hat{Y}_j(1-W_j) = Y_j(W_j) + \hat{\tau}_j \cdot (1_{\{w>W_j\}} - 1_{\{w<W_j\}})$, and sets the price to maximize predicted profit: $w_j^* = \arg\max_w (w \cdot 0.75 + (1-w) \cdot 0.25 - c) \cdot \hat{Y}_j(w)$. Hence, profit depends on which price is observed, W_j, even when $\hat{\tau}_j$ does not, as is the case for the "Infeasible: OLS" and "Diff. in Hyp. Choice" methods. For these two methods, we show profit based on applying $\hat{\tau}_j$ to observed prices from the randomized design, mirroring the other "Conventional Methods."

D.7 Estimation under endogenous treatment assignment

We select a virtual price for each product based on respondents' hypothetical willingness to pay (WTP) for it, which is correlated with potential outcomes.[52] Specifically, we set

$$W_{js} = 1\{\text{WTP}_j > \epsilon_{js}\},$$

for item j in simulation s, where random shocks ϵ_{js} are independent draws from a t-distribution.[53] We drop the observations of the real choices at the other price. This procedure simulates an environment in which sellers use consumer surveys to assess the attractiveness of their products when setting prices.[54] Because the analyst typically would not have access to those surveys, we do not include hypothetical WTPs in the vector $\boldsymbol{H}_j$.

Table A3 reports estimates of treatment effects using various methods when treatment assignment is endogenous. Column (1) repeats the ground truth estimate, that increasing the price from \$0.25 to \$0.75 changes the proportion of subjects buying the average snack by -0.075 percentage points.

The next two columns display estimates of treatment effects derived from regressions of the outcome on the treatment that control for conventional covariates, but do not otherwise address endogeneity. Column (2) reports an OLS regression. To allow for nonlinearities, we

[52] Appendix Figure A8 shows there is predictive relationship between each snack's actual purchase frequencies (potential outcomes) and the simulated probability it is observed at the high price. Alternative assignment mechanisms yield qualitatively similar conclusions.

[53] We set the mean of this distribution to the median of WTP, and set the standard deviation to that of the WTP distribution. We choose a fat-tailed distribution with 3 degrees of freedom so that even snacks with extreme WTPs still have a reasonable (if small) chance of being observed at either price. We draw 10,001 simulations, using an odd number so that the median is well-defined.

[54] We gathered data on real choices and hypothetical evaluations by drawing multiple samples from the same population. The respondents who provided the WTP data answered only one other hypothetical question. When we drop responses to that question from the specification, the only difference is that the estimate in Column (7) of Table A3 is -0.078 instead of -0.077.

also use approximate residual balancing (ARB, Athey, Imbens, and Wager 2018) with the same covariates as well as second-order terms and interactions (Column (3)).

For our method, we show results based on the same specifications of the prediction model as in Table 2. 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. For Columns (6), (7), and (8), we include higher order terms as well as more detailed hypothetical evaluations and estimate using ARB as described in Appendix C.3.

Controlling for conventional covariates in a regression of the outcome on the treatment (Columns (2) and (3)) yields estimates in the neighborhood of -0.03. In contrast, the multiple-covariate versions of our method yield estimates between -0.070 and -0.081. The most accurate specifications include the four basic hypothetical choice variables with 2nd order terms and interactions, as well as physical characteristics. Overall, estimates are quite stable across specifications. We offer some formal explanations for this favorable performance in Section 6.

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.

F An explicit model of underlying processes

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

F.1 Treatments and choices

We consider applications with settings (indexed $j = 1, \ldots, J$, representing treatment units such as goods, geographical jurisdictions, or markets) in which a set of individuals (indexed i) make choices, Y_{ij}, subject to the treatment assigned to that setting, $W_j \in \mathbb{W}$. The set of individuals may be identical across settings, overlapping between settings, or disjoint.[55]

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

Individual i's choice in setting j depends on the treatment, stable characteristics of the setting, $\boldsymbol{X}_j$ and $\boldsymbol{\eta}_j$, and unobserved *realized* conditions $\boldsymbol{\xi}_{ij} \sim F_j$ that i experiences in setting j. Thus, $Y_{ij} = Y(W_j, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$.[56] We are primarily concerned with either binary choices $Y_{ij} \in \{0, 1\}$ or continuous choices $Y_{ij} \in \mathbb{R}$.

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

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

[56] 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.

and there is a relationship between the distribution F_j^{typ} that affects treatment and the distribution F_j that affects choices.

The average outcome in setting j with treatment state w is

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

under typical conditions, and is

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

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

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

F.2 Motivations

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

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

We assume that these motivations reflect the treatment as well as the observed and unobserved characteristics of the individual and the setting: $\boldsymbol{\theta}_{ij}(w) = \boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$ or $\boldsymbol{\theta}_{ij}(w) = \boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij}^{typ})$, depending on whether the motivations are formed under actual or under typical conditions. At this level of generality, external conditions, including the treatment, affect choices only indirectly through motivations. This exclusion restriction should not be controversial, inasmuch as choices are governed by internal representations of decision problems. It follows that

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

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

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

Assume for the moment that we observe the potential outcomes $Y_j^{typ}(w)$ under typical conditions in *both* treatment states. Suppose we regress $Y_j^{typ}(w)$ on the distributional characteristics $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(w)$, pooling observations from all settings and treatment conditions, and then use the estimated equations to compute fitted choices, $\hat{Y}_j(0)$ and $\hat{Y}_j(1)$. As long as we select a functional specification with sufficient flexibility to accommodate the variation in conditional expectations, the treatment effect under typical conditions, $Y_j^{typ}(1) - Y_j^{typ}(0)$, will equal the fitted treatment effect, $\hat{Y}_j(1) - \hat{Y}_j(0)$.[57]

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

When we observe data only for the actual treatment states, those observations are systematically selected. However, by assumption, the treatment depends only on the features of the setting and typical conditions $(\boldsymbol{X}_j, \boldsymbol{\eta}_j, F_j^{typ})$. Because these factors affect outcomes only through $\boldsymbol{\theta}_{ij}(W_j)$, which we have assumed is observed, the treatment is unconfounded. It follows that observing just one of the potential outcomes for each setting does not cause systematic biases. Formally, the covariates $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(0)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(1)$ are *balancing scores* in the sense of Rosenbaum and Rubin (1983).

The other difference between our procedure and the (infeasible) fitted treatment effect procedure is that we use data on $Y_j(W_j)$ rather than $Y_j^{typ}(W_j)$. However, we will still correctly estimate the relationship between $Y_j^{typ}(W_j)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$ as long as the differences between (average) outcomes under realized and typical conditions, $\epsilon_j(W_j)$, are not systematically related to the distributions of typical intentions $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$. This assumption is plausible if the difference reflects sampling, or if conditions modulate baseline intentions

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

(and hence outcomes) in a similar way across settings. It is particularly natural for cases involving linear relationships between choices and measured intentions: if $\epsilon_j(W_j)$ and $\boldsymbol{D}_j^{typ,\boldsymbol{\theta}}(W_j)$ were correlated, then presumably F_j^{typ} would not reflect the most representative conditions.

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

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