Online Appendices

A Additional figures and tables cited in the text

Figure A1: Snack Demand Application: A Typical Choice Task

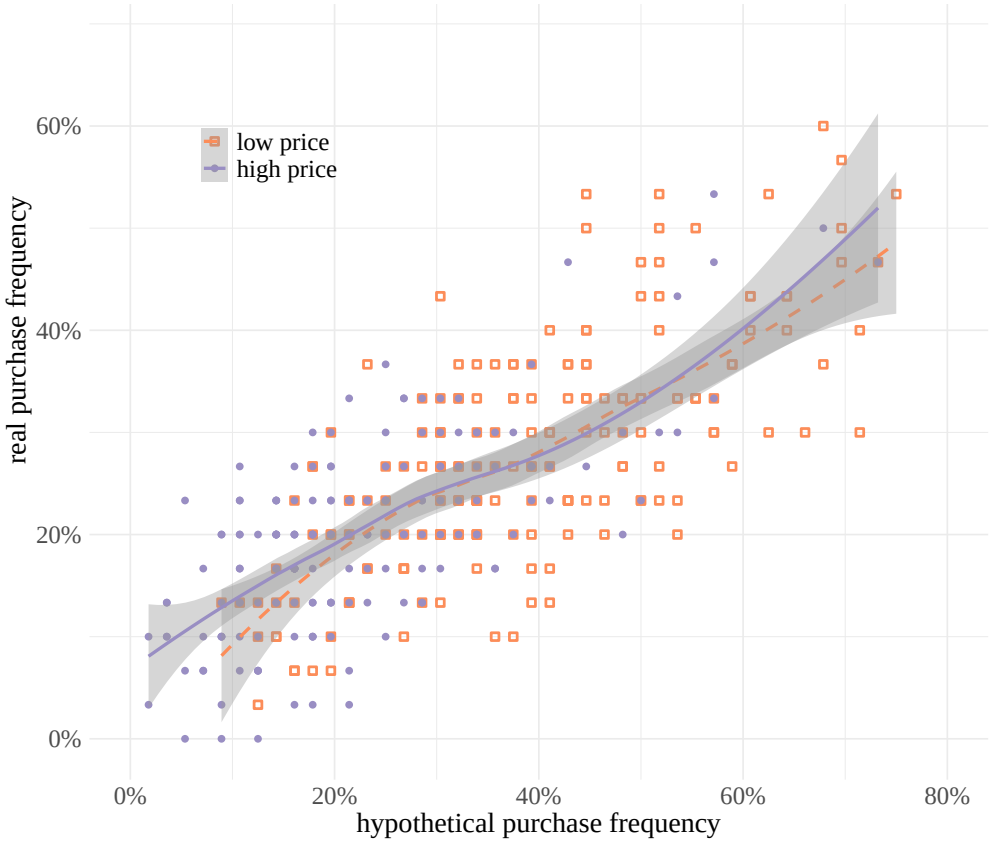


Figure A2: Real vs. Hypothetical Choices (nonparametric) in Snack Demand Application
Item-price pairs plotted. The lines and pointwise 95% confidence intervals shown are based on local quadratic regressions using the `R` function `stats::loess` with default parameters. Estimates at each point use 75% of the data nearest to the point with tricubic weighting.

Table A1: Regression Estimates in First Step in Microfinance Application

	Our Method: Low Dimensional			
	(5)	(6)	(7)	(8)
projected quintile	0.94	-0.17	-37.12	-26.90
	(1.64)	(1.57)	(17.44)	(16.60)
projected quintile, sq.			6.25	4.43
			(2.84)	(2.70)
own lending likelihood	-2.04	-0.23	3.46	0.62
	(1.27)	(1.21)	(11.86)	(11.15)
own lending likelihood, sq.			-0.65	-0.10
			(1.39)	(1.31)
other lending likelihood	2.77	1.25	1.60	5.49
	(1.55)	(1.46)	(15.80)	(14.90)
other lending likelihood, sq.			0.21	-0.44
			(1.81)	(1.71)
included in top 10	-3.23	3.34	5.94	4.35
	(2.46)	(2.57)	(8.75)	(8.29)
included in top 10, sq.			-14.56	-2.36
			(12.60)	(12.02)
Intercept	-0.06	2.82	45.55	30.77
	(2.39)	(2.31)	(14.12)	(13.75)
Gender = male		-1.63		-1.50
		(0.47)		(0.47)
Region = Asia		-1.38		-1.44
		(0.62)		(0.62)
Region = Africa		-1.34		-1.39
		(0.57)		(0.57)
Region = Central America		-1.99		-2.11
		(0.76)		(0.76)
Sector = Agriculture		-0.99		-0.75
		(0.51)		(0.51)
Sector = Food		-0.90		-0.52
		(0.63)		(0.64)
Sector = Housing		-3.88		-3.54
		(0.66)		(0.68)
Sector = Retail		0.38		0.45
		(0.75)		(0.74)
Sample size (outcome)	300	300	300	300

Estimated models in the first step of our method for the effect of matching on the inverse hyperbolic sine of fundraising velocity, within the first day. Column numbers correspond to the specifications of Table 1. Controls include dummies for gender, region, and sector. 'projected quintile' refers to the mean responses concerning projected quintile for fundraising velocity, 'own lending likelihood' to the contribution likelihood of the respondent, 'other lending likelihood' to the contribution likelihood of a typical user, and 'included in top 10' to the fraction of respondents who included the posting in their funding allocation task. Rows labeled 'sq.' refer to coefficients on the quadratic terms for the mean responses and frequencies of each hypothetical response. Analytical standard errors in parentheses.

Table A2: Estimated Treatment Effects from Microfinance Application With Sample Selection Correction

	Hypotheticals as Predictors			
	Low dimensional			
	(5)	(6)	(7)	(8)
Estimated effect of matching	0.52	0.37	1.69	1.31
	[0.51]	[0.39]	[0.49]	[0.46]
Sample selection correction	X	X	X	X
Controls		X		X
Hypotheticals:				
...avg. hypothetical eval.	X	X	X	X
...2nd order			X	X
Sample size	300	300	300	300
Observed matched		endogenous		
Observed unmatched		endogenous		

Estimates of the effect of matching on the inverse hyperbolic sine of fundraising velocity, within the first day. The sample selection correction uses the estimator $\hat{\tau}^{\text{hypit}}$. 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. '2nd order' includes quadratic terms for the mean responses. Bootstrap standard errors in square brackets.

Table A3: 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		
Test: = ground truth, p-value	1	0.42	0.86	0.60	0.67	0.48
(adjusted)	(1)	(0.94)	(0.98)	(0.96)	(0.96)	(0.96)
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		
Test: = ground truth, p-value	1	0.00	0.00	0.01	0.06	0.90
(adjusted)	(1)	(0.00)	(0.04)	(0.10)	(0.32)	(0.90)
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. Adjusted p-values apply a stepdown procedure to adjust for multiple hypothesis testing for the tests in all panels of this table (Romano and Wolf 2010; List, Shaikh, and Vayalinkal 2023). For results excluding WTP see Table 3.

Table A4: Estimating Snack 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)
Test: = ground truth, p-value	1	0.00	0.00	0.00	0.00	0.00
(adjusted)	(1)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)
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)
Test: = ground truth, p-value	1	0.33	0.34	0.90	0.00	0.80
(adjusted)	(1)	(0.73)	(0.69)	(0.90)	(0.02)	(0.95)
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)
Test: = ground truth, p-value	1	0.16	0.00	0.52	0.00	0.00
(adjusted)	(1)	(0.48)	(0.02)	(0.83)	(0.01)	(0.00)
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. Adjusted p-values apply a stepdown procedure to adjust for multiple hypothesis testing for the tests in all panels of this table (Romano and Wolf 2010; List, Shaikh, and Vayalinkal 2023).

Table A5: Regression Estimates in First Step in Snack Demand Application (multivariate)

	Low Price		High Price	
	(2)	(3)	(2)	(3)
hypothetical choice	0.30	0.29	0.44	0.39
	(0.08)	(0.08)	(0.08)	(0.08)
cheap talk	0.23	0.16	-0.04	-0.07
	(0.07)	(0.07)	(0.07)	(0.07)
intensity	-0.03	-0.03	0.01	0.03
	(0.08)	(0.07)	(0.07)	(0.07)
vicarious choice	0.02	0.11	0.14	0.17
	(0.06)	(0.06)	(0.05)	(0.06)
WTP as choice	0.12	0.12	0.02	0.02
	(0.08)	(0.07)	(0.04)	(0.04)
Intercept	-0.01	-0.05	0.03	-0.02
	(0.04)	(0.05)	(0.02)	(0.03)
Calories per gram of serving		0.01		0.01
		(0.03)		(0.02)
Calories from fat per gram of serving		-0.00		-0.02
		(0.03)		(0.02)
Fat content per gram of serving		0.02		0.17
		(0.28)		(0.23)
Sodium per gram of serving		0.00		0.00
		(0.00)		(0.00)
Carbohydrates per gram of serving		-0.07		-0.06
		(0.11)		(0.09)
Sugar per gram of serving		-0.04		0.01
		(0.05)		(0.04)
Protein per gram of serving		-0.07		-0.04
		(0.16)		(0.13)
Serving size in grams		0.00		0.00
		(0.00)		(0.00)
Sample size (outcome)	189	189	189	189
Test: = low price (p-value)			0.00	0.03

Estimated models in the first step of our method for the effect the high price (vs. low price) on the real purchase frequency, corresponding to the specifications in Table 3. The final row reports a test of equality of all coefficients in a given specification (including intercept and controls) when estimated on the low-price and high-price samples, with p-value adjusted for multiple hypothesis testing within table.

Table A6: Regression Estimates in First Step in Snack Demand Application (univariate)

	Low Price					High Price				
	(2)	(3)	(4)	(5)	(6)	(2)	(3)	(4)	(5)	(6)
hypothetical choice	0.53					0.52				
	(0.04)					(0.04)				
cheap talk		0.54					0.46			
		(0.04)					(0.05)			
intensity as choice			0.54					0.59		
			(0.05)					(0.07)		
vicarious as choice				0.53					0.54	
				(0.05)					(0.05)	
WTP as choice					0.55					0.29
					(0.06)					(0.04)
Intercept	0.07	0.08	0.15	0.14	-0.15	0.08	0.12	0.13	0.11	0.05
	(0.02)	(0.02)	(0.01)	(0.02)	(0.04)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Sample size (outcome)	189	189	189	189	189	189	189	189	189	189
Test: = low price (p-value)						0.50	0.02	0.66	0.01	0.00

Estimated models in the first step of our method for the effect the high price (vs. low price) on the real purchase frequency, corresponding to the specifications in Appendix Table A4b. Analytical standard errors in parentheses. The final row reports a test of equality of coefficient and intercept in a given specification when estimated on the low-price and high-price samples, with p-value adjusted for multiple hypothesis testing within table.

Table A7: 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 A8: Snack Demand Treatment Effects: Endogenous Prices Controlling for Hypotheticals

	Ground Truth	Observational			
	Experiment (1)	OLS (2)	OLS (3)	ARB (4)	ARB (5)
Median estimated effect of high price	-0.075	-0.072	-0.070	-0.069	-0.073
Median standard error	(0.004)	(0.011)	(0.014)	(0.012)	(0.013)
Controls		X	X	X	X
Hypotheticals (of both treatment states):					
... all hypothetical choices (excl. WTP)		X	X	X	X
... detailed hypothetical eval. (excl. WTP)					X
2nd order + interactions				X	X
interaction with treatment			X	X	X
Sample size (outcome)	189 ($\times 2$)	189	189	189	189
Observed at high price	All	$\text{WTP}_j > \epsilon_{js}$			
Observed at low price	All	$\text{WTP}_j \leq \epsilon_{js}$			

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.

B A formal justification for Assumption 2.3.1 in the presence of treatment endogeneity

In this section, we provide an explicit model of a plausible underlying process for which treatment assignment is unconfounded conditional on hypothetical evaluations (Assumption 2.3.1), even though treatment is endogenous and hypothetical evaluations do not capture all the conditions contributing to the choice aggregate in each setting. It is worth emphasizing that we intend this model only as an illustration of the types of processes for which Assumption 2.3.1 might hold despite treatment endogeneity. To be clear, the purpose of this illustration is to show that treatment endogeneity does not *automatically* overturn Assumption 2.3.1. We do not mean to suggest that Assumption 2.3.1 will *generally* hold when treatment is endogenous.

B.1 The model

As in the main text, 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 $i = 1, ...I$) make choices, Y_{ij}, subject to the treatment assigned to that setting, $W_j \in \mathbb{W}$. Individual i's choice in setting j depends on the psychological *motivations,* $\boldsymbol{\theta}_{ij}(w)$, that arise under treatment state w:

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

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.

Psychological motivations depend on treatment, w, the setting's stable characteristics $\boldsymbol{X}_j$ and $\boldsymbol{\eta}_j$, which are respectively observable and unobservable to the econometrician, and unobserved conditions $\boldsymbol{\xi}_{ij}$ that i experiences in setting j. Thus, $\boldsymbol{\theta}_{ij}(w) = \boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij})$ and $Y_{ij}(w) = Y(\boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij}))$. Within setting j, the distribution of $\boldsymbol{\xi}_{ij}$ across individuals is governed by a CDF, $F_j(\cdot \mid \upsilon_j)$, where υ_j represents transient features of the setting. The distribution of υ_j is governed by a CDF, G_j. The average outcome in setting j with treatment state w, conditional on υ_j, is therefore

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

When people answer hypothetical questions about their choices w, they do not necessarily envision the specific conditions, $\boldsymbol{\xi}_{ij}$, that will materialize when the treatment is actually in effect. Instead, we assume they respond based on typical conditions, which we write as $\boldsymbol{\xi}_{ij}^{typ}$. Within setting j, the distribution of $\boldsymbol{\xi}_{ij}^{typ}$ across individuals is governed by a CDF, F_j^{typ}. We assume that $F_j^{typ} = F_j(\cdot \mid \upsilon_j^{typ})$, where υ_j^{typ} represents the typical features of setting j, rather than the features that actually materialize. Accordingly, the average choice in setting j with treatment state w under typical conditions υ_j^{typ} is

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

We can therefore write

$$Y_j(w) = Y_j^{typ}(w) + \epsilon_j(w),$$

The error term $\epsilon_j(w)$ reflects the difference between distributions $F_j(\cdot \mid \upsilon_j)$ and $F_j(\cdot \mid \upsilon_j^{typ})$, which is in turn attributable to the difference between the realized features of setting j, υ_j, and setting j's typical features, υ_j^{typ}.

We assume that the treatment assigned to setting j may depend on the setting's stable characteristics $\boldsymbol{X}_j$ and $\boldsymbol{\eta}_j$, as well as the distribution of typical conditions facing individuals, $F_j(\cdot \mid \upsilon_j^{typ})$. Thus, $W_j = W_j(\boldsymbol{X}_j, \boldsymbol{\eta}_j, \upsilon_j^{typ})$. Notice that this assumption covers the case in which the treatment depends on the average choice for the setting under typical conditions, $Y_j^{typ}(w)$. For example, the treatment selector might attempt to assess $Y_j^{typ}(w)$ in advance through surveys resembling our hypothetical choice elicitations. In contrast, our assumption rules out the possibility that treatment assignment depends on the difference in the average choice under realized and typical conditions, $Y_j(w) - Y_j^{typ}(w) = \epsilon_j(w)$. Our justification for this restriction is that treatment selection occurs before the transient features of the setting, υ_j, are realized.[53]

Treatment endogeneity may arise from two sources. First, unobservable factors $\boldsymbol{\eta}_j$ affect both treatment and choices. Second, some components of $\boldsymbol{\upsilon}_j^{typ}$ may also be unobserved; they affect treatment and are related (across settings) to the realizations $\boldsymbol{\upsilon}_j$ that directly impact choices. However, under our assumptions, the information contained in $\epsilon_j(w)$ is not

[53] In principle, the actor choosing the treatment might consider G_j, the entire distribution of υ_j, rather than just υ_j^{typ}. 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.

available at the time of treatment selection. Hence, it is natural to assume that this error is orthogonal to treatment, conditional on the determinants of typical choices:

$$W_j \perp\!\!\!\perp \epsilon_j(w) \mid \left(\theta_{ij}^{typ}(w)\right)_{i\in\{1,\ldots,I\}}.$$

To put it another way, an ability to predict $\epsilon_j(W_j)$ based on the conditions leading to W_j would imply that $\left(\theta_{ij}^{typ}(w)\right)_{i\in\{1,\ldots,I\}}$ does not reflect the most representative conditions.

B.2 Estimation and unconfoundedness

Due to hypothetical bias, hypothetical choice data does not yield accurate measures of $Y_j^{typ}(w)$. For the moment, assume it is possible to assess $\boldsymbol{\theta}_{ij}^{typ}(w) = \boldsymbol{\theta}(w, \boldsymbol{X}_j, \boldsymbol{\eta}_j, \boldsymbol{\xi}_{ij}^{typ})$ accurately through hypothetical questions. Let $\boldsymbol{D}_j^{typ}(w)$ be a vector of statistics describing features of the marginal distribution of these measures, such as moments and percentiles, for setting j. If $\boldsymbol{D}_j^{typ}(w)$ includes a sufficient collection of moments, we can express $Y_j^{typ}(w)$ as a function of $\boldsymbol{D}_j^{typ}(w)$. For simplicity, we take this function to be linear and assume the treatment is binary.

Assume for the moment that we observe the typical outcomes $Y_j^{typ}(w)$, rather than the actual outcomes $Y_j(w)$, under typical conditions in *both* treatment states. Suppose we regress $Y_j^{typ}(w)$ on the distributional characteristics $\boldsymbol{D}_j^{typ}(w)$, pooling observations from all settings and treatment conditions, and then use the estimated equations to compute fitted choices, $\hat{Y}_j(0)$ and $\hat{Y}_j(1)$. Under our assumptions, the treatment effect for typical conditions, $Y_j^{typ}(1) - Y_j^{typ}(0)$, will equal the fitted treatment effect, $\hat{Y}_j(1) - \hat{Y}_j(0)$.

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 $\boldsymbol{D}_j^{typ}(W_j)$), 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.

When we observe data only for the actual treatment states, W_j, those observations are systematically selected. However, by assumption, the treatment depends only on the features of the setting and typical conditions $(\boldsymbol{X}_j, \boldsymbol{\eta}_j, \upsilon_j^{typ})$. Because these factors affect outcomes only through $\boldsymbol{\theta}_{ij}(W_j)$, which we have assumed is observed, $Y_j^{typ}(W_j)$ does not vary with W_j conditional on $\boldsymbol{D}_j^{typ}(W_j)$. Furthermore, we have posited that $\epsilon_j(W_j)$ is orthogonal to W_j conditional on $\left(\theta_{ij}^{typ}(w)\right)_{i\in\{1,\ldots,I\}}$. Therefore, treatment assignment is (conditionally)

unconfounded, and regressing $Y_j(W_j)$ on $\boldsymbol{D}_j^{typ}(W_j)$ should correctly identify the relationship between $Y_j(w)$ on $\boldsymbol{D}_j^{typ}(w)$.

While motivations are necessarily measured imperfectly, that limitation is not necessarily problematic. The same reasoning applies as long as it is possible to elicit motive-specific hypothetical responses H_{ij} spanning θ_{ij} such that the relationship between θ_{ij} and those measures is stable. As noted in the main text, systematic misreporting of motives does not necessarily undermine the stability of that relationship.

C Proofs and additional econometric results

C.1 Proof of Proposition 1

The data are a random sample of independent observations $(Y_j, W_j, \boldsymbol{H}_j(W_j), \boldsymbol{H}_j(\underline{w}), \boldsymbol{H}_j(\overline{w}), \boldsymbol{X}_j)_{j=1}^J$ where $Y_j \in \mathbb{R}$, $W_j \in \mathbb{W}$, and $\boldsymbol{H}_j(w) \in \mathbb{R}^{Q_H}$ 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 $\hat{\tau} = \frac{1}{J}\sum_{j=1}^J (\boldsymbol{H}_j(\overline{w}) - \boldsymbol{H}_j(\underline{w}))\hat{\boldsymbol{\beta}}$.

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

$$g(y, \boldsymbol{z}_{\underline{w}}, \boldsymbol{z}_{\overline{w}}, \boldsymbol{z}, \tau, \boldsymbol{\delta}) = \tau - (\boldsymbol{z}_{\overline{w}} - \boldsymbol{z}_{\underline{w}})\boldsymbol{\delta}$$
$$\boldsymbol{m}(y, \boldsymbol{z}_{\underline{w}}, \boldsymbol{z}_{\overline{w}}, \boldsymbol{z}, \tau, \boldsymbol{\delta}) = \boldsymbol{z}^T(y - \boldsymbol{z}\boldsymbol{\delta})$$

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

$$\mathbb{E}(g(Y, \boldsymbol{Z}(\underline{w}), \boldsymbol{Z}(\overline{w}), \boldsymbol{Z}, \tau^*, \boldsymbol{\delta}^*) = 0$$

where $\tau^* = \mathbb{E}(Y(\overline{w}) - Y(\underline{w}))$ 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 or 3B), by the law of iterated expectations:

$$\mathbb{E}(\boldsymbol{m}(Y, \boldsymbol{Z}(\underline{w}), \boldsymbol{Z}(\overline{w}), \boldsymbol{Z}, \tau^*, \boldsymbol{\delta}^*) = \boldsymbol{0}_{Q\times 1}$$

where $\boldsymbol{0}_{Q\times 1}$ 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

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

and

$$\begin{aligned}\boldsymbol{\Psi} &= \mathbb{E}(\boldsymbol{\psi}\boldsymbol{\psi}') = \mathbb{E}\left(\begin{bmatrix} g^2 & g\boldsymbol{m}^T \\ g\boldsymbol{m} & \boldsymbol{m}\boldsymbol{m}^T \end{bmatrix}\right) \\ &= \mathbb{E}\left(\begin{bmatrix} (\tau^* - (\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w}))\boldsymbol{\delta}^*)^2 & \boldsymbol{Z}(\tau^* - (\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w}))\boldsymbol{\delta}^*)(Y - \boldsymbol{Z}\boldsymbol{\delta}^*) \\ \boldsymbol{Z}^T(\tau^* - (\boldsymbol{Z}(\overline{w}) - \boldsymbol{Z}(\underline{w}))\boldsymbol{\delta}^*)(Y - \boldsymbol{Z}\boldsymbol{\delta}^*) & \boldsymbol{Z}^T\boldsymbol{Z}(Y - \boldsymbol{Z}\boldsymbol{\delta}^*)^2 \end{bmatrix}\right)\end{aligned}$$

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 $\boldsymbol{V}^{\text{ols}} = \mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}\Big)^{-1}\mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}(y - \boldsymbol{Z}\boldsymbol{\delta}^*)^2\Big)\mathbb{E}\Big(\boldsymbol{Z}^T\boldsymbol{Z}\Big)^{-1}$ 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

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

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. *Let $\mathbb{Z}_{\underline{w}}$ and $\mathbb{Z}_{\overline{w}}$ be the supports of the distributions of predictors in the control and treatment states, respectively. There exists $\eta > 0$ such that, for each $\boldsymbol{z} \in (\mathbb{Z}_{\underline{w}} \cup \mathbb{Z}_{\overline{w}})$, there is some $w \in \mathbb{W}$ with*

$$\Pr(W = w \mid \boldsymbol{Z}(w) = \boldsymbol{z}) > \eta.$$

Evaluations overlap states that, for any value of the predictors $\boldsymbol{z} \in (\mathbb{Z}_{\underline{w}} \cup \mathbb{Z}_{\overline{w}})$, 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 3B).

With this assumption we can ensure identification:

Proposition 4. *The average effect of the treatment, $\tau = \mathbb{E}(Y(1) - Y(0))$, is semiparametrically identified under Assumptions 1; one of 3A or 3B; 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 $w \in \mathbb{W}$. By the law of iterated expectations, $\mathbb{E}(Y(\overline{w})) = \mathbb{E}(\mathbb{E}(Y_j(\overline{w}) \mid \boldsymbol{Z}_j(\overline{w})))$. By the evaluations overlap assumption, for any $\boldsymbol{z} \in \mathbb{Z}_{\overline{w}}$, there is some $w \in W$ with $\Pr(W = w \mid \boldsymbol{Z}(w) = \boldsymbol{z}) > 0$. Hence, $\mathbb{E}(Y(w) \mid \boldsymbol{Z}(w) = \boldsymbol{z}, W = w)$ is identified by $\mathbb{E}(Y \mid \boldsymbol{Z}(w) = \boldsymbol{z}, W = w)$ for some $w \in \mathbb{W}$. By unconfoundedness, $\mathbb{E}(Y(w) \mid \boldsymbol{Z}(w) = \boldsymbol{z}, W = w) = \mathbb{E}(Y(w) \mid \boldsymbol{Z}(w) = \boldsymbol{z})$. By mapping invariance, $\mathbb{E}(Y(w) \mid \boldsymbol{Z}(w) = \boldsymbol{z}) = \mathbb{E}(Y(\overline{w}) \mid \boldsymbol{Z}(\overline{w}) = \boldsymbol{z})$. Hence, $\mathbb{E}(Y(\overline{w}))$ is identified. The argument for $\mathbb{E}(Y(\underline{w}))$ is similar. Hence, $\tau_{\overline{w},\underline{w}}$ 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 Measurement Error

Suppose the Invariant Mapping assumption holds with *population* average hypothetical evaluations $\boldsymbol{H}_j(w)$ for each setting. In practice, the researcher observes evaluations of setting j under treatment state w from a random sample $\mathcal{R}_j(w)$ of respondents indexed by i, $\boldsymbol{H}_{ij}(w)$. Respondents may evaluate multiple settings or treatment states, such that $\mathcal{R}_j(w) \cap \mathcal{R}_{j'}(w')$ need not be empty. Let the minimum number of responses per setting and treatment state be $r_1 = \min_{(j,w)} |\mathcal{R}_j(w)|$. Let the maximum number of settings and treatment states evaluated by a single respondent be $r_2 = \max_i |\{(j,w) : \ i \in \mathcal{R}_j(w)\}|$.

Proposition 1 derives the statistical properties of the estimator $\hat{\tau}$ when $\boldsymbol{H}_j(w)$ is observed. What are the statistical properties of estimators for $\tau_{\overline{w},\underline{w}}$ when only $\hat{\boldsymbol{H}}_j(w)$ is observed instead of $\boldsymbol{H}_j(w)$?

We provide three analyses: (i) if $r_1/\sqrt{J} \to \infty$ the estimator is $\sqrt{J}$ consistent despite measurement error, and if also $r_2/r_1 \to 0$, the measurement error does not affect the asymptotic distribution of the estimator; (ii) correlation across observations due to "respondent fixed effects" does not affect the estimator if each respondent evaluates every setting and treatment state even if $r_2/r_1 \not\to 0$; and (iii) if $r_1 > 1$ is small, one can use standard corrections for measurement error.

C.3.1 Rate conditions

The sample average of responses is $\hat{\boldsymbol{H}}_j(w) = \frac{1}{|\mathcal{R}_j(w)|} \sum_{i \in \mathcal{R}_j(w)} \boldsymbol{H}_{ij}(w)$. We investigate this question assuming $r_1 \to \infty$, albeit possibly slowly relative to the sample size J.

Under the assumptions of Proposition 1, the estimator $\hat{\tau}$ is a just-identified GMM estimator based on moments

$$g(y, \boldsymbol{z}_{\underline{w}}, \boldsymbol{z}_{\overline{w}}, \boldsymbol{z}, \tau, \boldsymbol{\delta}) = \tau - (\boldsymbol{z}_{\overline{w}} - \boldsymbol{z}_{\underline{w}})\boldsymbol{\delta}$$
$$\boldsymbol{m}(y, \boldsymbol{z}_{\underline{w}}, \boldsymbol{z}_{\overline{w}}, \boldsymbol{z}, \tau, \boldsymbol{\delta}) = \boldsymbol{z}^T(y - \boldsymbol{z}\boldsymbol{\delta})$$

satisfying $\mathbb{E}(g(Y, \boldsymbol{Z}(\underline{w}), \boldsymbol{Z}(\overline{w}), \boldsymbol{Z}, \tau^*_{\overline{w},\underline{w}}, \boldsymbol{\delta}^*) = 0$ and $\mathbb{E}(\boldsymbol{m}(Y, \boldsymbol{Z}(\underline{w}), \boldsymbol{Z}(\overline{w}), \boldsymbol{Z}, \tau^*_{\overline{w},\underline{w}}, \boldsymbol{\delta}^*) = \boldsymbol{0}_{Q\times 1}$, where $\tau^*_{\overline{w},\underline{w}} = \mathbb{E}(Y_j(\overline{w}) - Y_j(\underline{w}))$ and $\boldsymbol{\delta}^* \equiv [\boldsymbol{\beta}^T, \ \boldsymbol{\gamma}^T]^T$ with $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ defined in Assumption 2. Here, we use the shorthand $\boldsymbol{Z}_j(w) = [\boldsymbol{H}_j(w), \ \boldsymbol{X}_j]$ and $\boldsymbol{Z}_j = \boldsymbol{Z}_j(W_j)$. Let $\boldsymbol{k} = (g, \boldsymbol{m}^T)^T$.

Now consider the estimator $(\check{\tau}, \check{\boldsymbol{\delta}})$ using $\hat{\boldsymbol{Z}}_j(w) = [\hat{\boldsymbol{H}}_j(w), \ \boldsymbol{X}_j]$ and $\hat{\boldsymbol{Z}}_j = \hat{\boldsymbol{Z}}_j(W_j)$, which solves

$$\frac{1}{J}\sum_{j=1}^{J} \boldsymbol{k}(Y_j, \hat{\boldsymbol{Z}}_j(\underline{w}), \hat{\boldsymbol{Z}}_j(\overline{w}), \hat{\boldsymbol{Z}}_j, \check{\tau}, \check{\boldsymbol{\delta}}) = \boldsymbol{0}_{(1+Q)\times 1}.$$

By standard arguments, one can show that the $\sqrt{J}$-asymptotic behavior of $(\check{\tau}, \check{\boldsymbol{\delta}})$ around $(\tau^*_{\overline{w},\underline{w}}, \boldsymbol{\delta}^*)$ depends primarily on $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} \boldsymbol{k}(Y_j, \hat{\boldsymbol{Z}}_j(\underline{w}), \hat{\boldsymbol{Z}}_j(\overline{w}), \hat{\boldsymbol{Z}}_j, \tau^*_{\overline{w},\underline{w}}, \boldsymbol{\delta}^*)$. We consider the elements due to $\boldsymbol{m}$; g behaves similarly.

For $\boldsymbol{m}$, write

$$\frac{1}{\sqrt{J}}\sum_{j=1}^{J} \hat{\boldsymbol{Z}}_{\boldsymbol{j}}^{\ T}(Y_j - \hat{\boldsymbol{Z}}_j\boldsymbol{\delta}^*) = \frac{1}{\sqrt{J}}\sum_{j=1}^{J} \boldsymbol{Z}_j(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*) + \boldsymbol{Z}_j^T\boldsymbol{\eta}_j\boldsymbol{\delta}^* + \boldsymbol{\eta}_j^T(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*) - \boldsymbol{\eta}_j^T\boldsymbol{\eta}_j\boldsymbol{\delta}^*$$

where $\boldsymbol{\eta}_j = \frac{1}{|\mathcal{R}_j(w)|}\sum_{i\in\mathcal{R}_j(w)} \boldsymbol{\eta}_{ij}$ with $\boldsymbol{\eta}_{ij} = \hat{\boldsymbol{Z}}_{ij} - \boldsymbol{Z}_j$. The first term, $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} \boldsymbol{Z}_j(Y_j - \boldsymbol{Z}_j\boldsymbol{\delta}^*)$, is the same as when $\boldsymbol{H}_j(w)$ is observed. We derive conditions under which each element of the last three terms is small in probability. For notational simplicity, we suppress conditioning on $(\boldsymbol{Z}_j)_{j=1}^J$ in the expectations and variances below.

For the second term, $\mathbb{E}\Big(\frac{1}{\sqrt{J}}\sum_{j=1}^{J} \boldsymbol{Z}_j^T\boldsymbol{\eta}_j\boldsymbol{\delta}^*\Big) = \boldsymbol{0}_{Q\times 1}$ because respondents are a random sample such that $\mathbb{E}(\boldsymbol{\eta} \mid \boldsymbol{Z}) = \boldsymbol{0}_{1\times Q}$. Next, consider the variance of the l'th element of the second term,

$$\begin{aligned}\operatorname{var}\Big(\frac{1}{\sqrt{J}}\sum_{j=1}^{J} \boldsymbol{Z}_{j,l}\boldsymbol{\eta}_j\boldsymbol{\delta}^*\Big) =& \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{Z}_{j,l}^2\boldsymbol{\delta}^{*\prime}\mathbb{E}(\boldsymbol{\eta}_j^T\boldsymbol{\eta}_j)\boldsymbol{\delta}^* \\ &+ \frac{1}{J}\sum_{j=1}^{J}\sum_{j'=1}^{J} 1(j\neq j')\boldsymbol{Z}_{j,l}\boldsymbol{Z}_{j',l}\boldsymbol{\delta}^{*\prime}\mathbb{E}(\boldsymbol{\eta}_j^T\boldsymbol{\eta}_{j'})\boldsymbol{\delta}^*.\end{aligned}$$

The component $\mathbb{E}(\boldsymbol{\eta}_j^T\boldsymbol{\eta}_j)$ is the variance of the measurement error across respondents in $\mathcal{R}_j(W_j)$. Since these responses are independent of each other, the variance $\mathbb{E}(\boldsymbol{\eta}_j^T\boldsymbol{\eta}_j)$ is of order $|\mathcal{R}_j(w)|^{-1}$, and the term in the first line vanishes under standard regularity conditions as $\min_{j,w}|\mathcal{R}_j(w)| \to \infty$. The second line is due to respondents evaluating multiple settings. $\boldsymbol{\eta}_j$ consists of $|\mathcal{R}_j(W_j)|$ terms $\boldsymbol{\eta}_{ij}$, each receiving weight $|\mathcal{R}_j(W_j)|^{-1}$. For a given $\boldsymbol{\eta}_{ij}$, the covariance with $\boldsymbol{\eta}_{j'}$ is zero except when i also evaluated setting j' in treatment state $W_{j'}$, $i \in \mathcal{R}_{j'}(W_{j'})$, in which case the covariance is weighted by $|\mathcal{R}_{j'}(W_{j'})|^{-1}$. There are at most r_2 such terms. Hence, if $r_2/r_1 \to 0$, the second line vanishes in probability. The third term behaves similarly.

For the fourth term, $\mathbb{E}(\boldsymbol{\eta}_j^T\boldsymbol{\eta}_j)$ is of order $|\mathcal{R}_j(W_j)|^{-1}$. Hence, $\frac{1}{\sqrt{J}}\sum_{j=1}^{J}\mathbb{E}(\boldsymbol{\eta}_j^T\boldsymbol{\eta}_j)\boldsymbol{\delta}^*$ vanishes if the number of respondents per setting grows faster than the square root of the sample

size, $r_1/\sqrt{J} \to \infty$. Next, consider the variance of the l'th element of the fourth term,

$$\begin{aligned}\operatorname{var}\Big(\frac{1}{\sqrt{J}}\sum_{j=1}^{J}\boldsymbol{\eta}_{j,l}\boldsymbol{\eta}_j\boldsymbol{\delta}^*\Big) =& \frac{1}{J}\sum_{j=1}^{J}\operatorname{var}\Big(\boldsymbol{\eta}_{j,l}\boldsymbol{\eta}_j\Big)\boldsymbol{\delta}^* \\ &+ \frac{1}{J}\sum_{j=1}^{J}\sum_{j'=1}^{J}1(j\neq j')\boldsymbol{\delta}^{*\prime}\operatorname{cov}\Big(\eta_{j,l}\boldsymbol{\eta}_j, \eta_{j',l}\boldsymbol{\eta}_{j'}\Big)\boldsymbol{\delta}^*.\end{aligned}$$

The leading terms of the elements of the variance matrix $\operatorname{var}\Big(\boldsymbol{\eta}_{j,l}\boldsymbol{\eta}_j\Big)$ are of order $|\mathcal{R}_{j'}(W_{j'})|^{-2}$. Hence, the first line vanishees as $r_1 \to \infty$. Next, consider the (l', l'') element of the covariance matrix in the second line:

$$\frac{1}{|\mathcal{R}_j(W_j)|^2|\mathcal{R}_{j'}(W_{j'})|^2}\sum_{i\in\mathcal{R}_j(W_j)}\sum_{i'\in\mathcal{R}_j(W_j)}\sum_{i''\in\mathcal{R}_{j'}(W_{j'})}\sum_{i'''\in\mathcal{R}_{j'}(W_{j'})}\operatorname{cov}\Big(\eta_{ij,l}\eta_{i'j,l'}, \eta_{i''j',l}\eta_{i'''j',l''}\Big).$$

Fix j. By random sampling, the covariance is zero whenever a respondent index appears only in a single argument. If $i \neq i'$, the covariance is zero unless $i = i''$ and $i' = i'''$ (or $i = i'''$ and $i' = i''$). By definition of r_2, $i = i''$ at most r_2 times. Hence, the total contribution (across j') of $i \neq i'$ terms for a given i and i' is, in the worst case, of order $O(\frac{r_2}{|\mathcal{R}_j(W_j)|^2|\mathcal{R}_{j'}(W_{j'})|^2})$. There are of order $|\mathcal{R}_j(W_j)|^2$ many $i \neq i'$ terms, so the total contribution of such terms is of order $\frac{r_2}{|\mathcal{R}_{j'}(W_{j'})|^2}$ at most. If $i = i'$, the covariance is zero unless $i = i' = i'' = i'''$. For fixed j, there are at most $|\mathcal{R}_j(W_j)|r_2$ such terms, so their total contribution is of order at most $\frac{r_2}{|\mathcal{R}_j(W_j)||\mathcal{R}_{j'}(W_{j'})|^2}$. Hence, the second line here vanishes faster than the term requiring $r_2/r_1 \to 0$ above.

Combining results, the estimator using $\hat{\boldsymbol{H}}_j(w)$ in place of $\boldsymbol{H}_j(w)$ converges at the usual rate if $r_1/\sqrt{J} \to \infty$ and is asymptotically equivalent to the estimator using $\boldsymbol{H}_j(w)$ if also $r_2/r_1 \to 0$.

C.3.2 Respondent fixed effects

Suppose responses are correlated within respondent across settings and treatment states due to "respondent fixed effects" $\boldsymbol{\eta}_{ij}(w) = \boldsymbol{\eta}_i + \boldsymbol{\nu}_{ij}(w)$ where $\boldsymbol{\nu}_{ij}(w)$ are independent across i, j, and w (but arbitrarily correlated across evaluations within i, j, and w). Also suppose every respondent i evaluates each setting and treatment state (j, w). Then $r_2/r_1 \to 0$ requires $r_1/J \to \infty$, which may be infeasible in practice. However, as we show below, correlation in measurement error across settings due to $\boldsymbol{\eta}_i$ is exactly orthogonal to the estimator $\hat{\tau}$ and hence does not affect its asymptotic distribution. Hence, following the derivations of rate

conditions in the general case, $r_1/\sqrt{J} \to \infty$ is sufficient even if $r_2/r_1 \not\to 0$.

When the same respondents evaluate all settings in the observed treatment states W_j, the respondent fixed effects $\boldsymbol{\eta}_i$ cause a constant (across settings j) shift to the covariates in the first step regression for $\hat{\tau}$. This covariate shift is absorbed in its entirety by the intercept term in the regression. The slope coefficients $\hat{\boldsymbol{\beta}}$ are numerically identical to their values without respondent fixed effects. In the second step for $\hat{\tau}$, the differences $\hat{\boldsymbol{H}}_j(\overline{w}) - \hat{\boldsymbol{H}}_j(\underline{w})$ also are numerically identical to their values without respondent fixed effects, for all j. Hence, respondent fixed effects have no effect on $\hat{\tau}$ at all under this setup.

C.3.3 Bias correction

In some setups, neither of the analyses above may be applicable because $r_1/\sqrt{J} \not\to \infty$. In such setups, corrections for measurement error may improve statistical properties of the resulting estimators. Specifically, the (few) responses $\boldsymbol{Z}_{ij}(W_j)$ serve as repeated measurements of $\boldsymbol{Z}_j(W_j)$, making standard corrections possible. For simplicity, suppose $|\mathcal{R}_j(W_j)| = 2$ for all j. In the first step linear regression for $\hat{\tau}$, one can then use one response for each setting as an instrument for the other response for the same setting to obtain consistent estimates of $\boldsymbol{\delta}$. In the second step for $\hat{\tau}$, measurement error averages out across settings j if r_2 is fixed. If $r_2 > 1$, one may wish to design the survey such that simple adjustments to standard errors such as clustering or the approach of Conley (1999) are applicable.

C.4 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 $w \in \mathbb{W}$. Let $\hat{\boldsymbol{\beta}}$ be the first step regression slope coefficients on $\boldsymbol{H}_j$.

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 $\boldsymbol{H}_j = \boldsymbol{H}_j(w) + (\boldsymbol{H}_j(W_j) - \boldsymbol{H}_j(w))$ and $\epsilon_j = \epsilon_j(w) + (\epsilon_j(W_j) - \epsilon_j(w))$. By linearity, $\operatorname{cov}(\boldsymbol{H}_j^T(w), \epsilon_j(w)) = 0$.

Hence,

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

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 $H_{j,l}(w)$ of $\boldsymbol{H}_j^T(w)$, $\operatorname{cov}(H_{j,l}(w), \epsilon_j(W_j) - \epsilon_j(w))^2 \leq \operatorname{var}(H_{j,l}(w)) \operatorname{var}(\epsilon_j(W_j) - \epsilon_j(w))$ and similarly for the other two terms. So

$$\begin{aligned}&|\operatorname{cov}(H_{j,l}(w), \epsilon_j(W_j) - \epsilon_j(w)))| \\ &\leq \operatorname{var}(H_{j,l}(w)) \sqrt{\frac{\operatorname{var}(\epsilon_j(W_j) - \epsilon_j(w)))}{\operatorname{var}(H_{j,l}(w))}},\end{aligned}$$

$$\begin{aligned}&|\operatorname{cov}(H_{j,l}(W_j) - H_j(w), \epsilon_j(w))| \\ &\leq \operatorname{var}(H_{j,l}(w)) \sqrt{\frac{\operatorname{var}(H_{j,l}(W_j) - H_{j,l}(w)))}{\operatorname{var}(H_{j,l}(w))}} \cdot \sqrt{\frac{\operatorname{var}(\epsilon_j(w))}{\operatorname{var}(H_{j,l}(w))}},\end{aligned}$$

and

$$\begin{aligned}&|\operatorname{cov}(H_{j,l}(W_j) - H_{j,l}(w), \epsilon_j(W_j) - \epsilon_j(w))| \\ &\leq \operatorname{var}(H_{j,l}(w)) \sqrt{\frac{\operatorname{var}(H_{j,l}(W_j) - H_{j,l}(w)))}{\operatorname{var}(H_{j,l}(w))}} \cdot \sqrt{\frac{\operatorname{var}(\epsilon_j(W_j) - \epsilon_j(w)))}{\operatorname{var}(H_{j,l}(w))}}\end{aligned}$$

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

$$\begin{aligned}\operatorname{var}(H_j) = &\operatorname{var}(H_j(w)) + \operatorname{var}(H_j(W_j) - H_j(w)) \\ &+ 2\operatorname{cov}(H_j(w), H_j(W_j) - H_j(w))\end{aligned}$$

So, using Cauchy-Schwarz to bound the covariance,

$$\begin{aligned}\operatorname{var}(H_j) \geq &\operatorname{var}(H_j(w)) + \operatorname{var}(H_j(W_j) - H_j(w)) \\ &- 2\sqrt{\operatorname{var}(H_j(w))}\sqrt{\operatorname{var}(H_j(W_j) - H_j(w))}.\end{aligned}$$

Factor out $\mathrm{var}(H_j(w))$ to get

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

Hence, we can bound the norm of $\frac{\mathrm{cov}(H_j,\epsilon_j)}{\mathrm{var}(H_j)}$ using the norm with the right-hand-side above in place of $\mathrm{var}(H_j)$ in the denominator. Then $\mathrm{var}(H_j(w))$ cancels in the worst-case bounds for numerator and denominator. Finally, under Assumption 3C, the ratios $\frac{\mathrm{var}(H_j(W_j)-H_j(w)))}{\mathrm{var}(H_j(w))} \to 0$, $\frac{\mathrm{var}(\epsilon_j(W_j)-\epsilon_j(w)))}{\mathrm{var}(H_j(w))} \to 0$, and $\sqrt{\frac{\mathrm{var}(\epsilon_j(w))}{\mathrm{var}(H_j(w))}} < \infty$, so each remaining component of the numerator (after canceling $\mathrm{var}(H_j(w))$ against the denominator) vanishes. In contrast, the remaining denominator,

$$1 + \frac{\mathrm{var}(H_j(W_j) - H_j(w))}{\mathrm{var}(H_j(w))} - 2\sqrt{\frac{\mathrm{var}(H_j(W_j) - H_j(w))}{\mathrm{var}(H_j(w))}} \to 1.$$

Hence, the ratio vanishes such that $\lim_{k\to\infty} \mathrm{plim}_{n\to\infty} \hat{\beta}_k - \beta_k = 0$.

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

C.5 Sample selection model

C.5.1 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 (3D). **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*

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

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 3D, 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 $\hat{\tau}^{\text{hypit}} = \frac{1}{J}\sum_{j=1}^{J}(\boldsymbol{H}_j(1) - \boldsymbol{H}_j(0))\hat{\boldsymbol{\beta}}$. 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.5.3.

Proposition 5. *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 3D, the parametric estimator* $\hat{\tau}^{hypit}$ *is consistent for the average treatment effect* τ *and is asymptotically normal:*

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

with a formula for $V_{\tau,hypit}$ *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. The estimates presented in Table A9 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 3D, we obtain 1.31 (0.46).

Table A9: Estimated Treatment Effects from Microfinance Application Controlling for Hypotheticals

	Ground Truth	Observational Methods			
	Experiment (IV) (1)	OLS (2)	OLS (3)	ARB (4)	ARB (5)
Estimated effect of matching	1.24	2.98	3.26	3.02	2.90
	(0.33)	(0.39)	(0.40)	(0.31)	(0.34)
	[0.32]	[0.37]	[0.39]	[0.30]	[0.36]
Test: = ground truth, p-value	1	0.00	0.00	0.00	0.00
(adjusted)	(1)	(0.00)	(0.00)	(0.00)	(0.00)
Controls		X	X	X	X
Hypotheticals (of both treatment states):					
. . . avg. hypothetical eval.		X	X	X	X
. . . freq. hypothetical eval.					X
. . . 2nd order				X	X
interactions				X	X
interaction with treatment			X	X	X
Sample size	1091	300	300	300	300
Observed matched	use randomized variation		endogenous		
Observed unmatched	use randomized variation		endogenous		

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

C.5.2 Proof of Proposition 5

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 estimate $\boldsymbol{\beta}$. 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 $V_{\text{hypit}} = \mathbb{E}(\nabla \boldsymbol{\psi}^*)^{-1}\mathbb{E}(\boldsymbol{\psi}^*)\mathbb{E}(\nabla \boldsymbol{\psi}^{*,T})^{-1}$ (Newey and McFadden 1994). The $(1,1)$ element of V_{hypit} is the asymptotic variance of the treatment effect estimator, $V_{\tau,\text{hypit}}$.

C.5.3 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.6 Partial identification

C.6.1 Semiparametric partial identification: Proof of Proposition 3

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 $((Y(w), \boldsymbol{H}(w))_{w\in\mathbb{W}}, W, \boldsymbol{X})$ that is consistent with the distribution of the observable $(Y, W, (\boldsymbol{H}(w))_{w\in\mathbb{W}}, \boldsymbol{X})$ and yields that point as the average treatment effect. In Appendix C.6.2, 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.

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

$\mathbb{E}(\mathbb{E}(Y(\overline{w}) \mid \boldsymbol{H}(\overline{w}), \boldsymbol{X})) - \mathbb{E}(\mathbb{E}(Y(\underline{w}) \mid \boldsymbol{H}(\underline{w}), \boldsymbol{X}))$. Writing the outer expectations in integral form and using mapping invariance, we can write the treatment effect as $\tau = \int\int \mu(\boldsymbol{h}, \boldsymbol{x})(f_{(\boldsymbol{H}(\overline{w}),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) - f_{(\boldsymbol{H}(\underline{w}),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}))\,\mathrm{d}\boldsymbol{h}\mathrm{d}\boldsymbol{x}$, where the densities of $(\boldsymbol{H}(\overline{w}), \boldsymbol{X})$ and $(\boldsymbol{H}(\underline{w}), \boldsymbol{X})$, $f_{(\boldsymbol{H}(\overline{w}),\boldsymbol{X})}$ and $f_{(\boldsymbol{H}(\underline{w}),\boldsymbol{X})}$, are point identified from the data, and $\mu(\boldsymbol{h}, \boldsymbol{x}) = \mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})$ for all $w \in \mathbb{W}$. By the law of total probability,

$$\begin{aligned}\mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}) = &\Pr(W = w \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})\mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}, W = w) \\ &+ \Pr(W \neq w \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})\mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}, W \neq w).\end{aligned}$$

The conditional probabilities and the conditional expectation $\mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}, W = w)$ are point identified from the data. We can bound the final conditional expectation using $y_{\text{lb}} \leq \mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}, W \neq w) \leq y_{\text{ub}}$. The identifying power of the mapping invariance assumption is that we need only consider the *intersection* of the resulting bounds on $\mathbb{E}(Y(w) \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x})$ for all $w \in \mathbb{W}$ as a bound for $\mu(\boldsymbol{h}, \boldsymbol{x})$, which can be noticeably tighter than the separate bounds.[54] From the integral equation above, the upper bound for $\tau_{\overline{w},\underline{w}}$ then uses the upper bound on μ whenever $f_{(\boldsymbol{H}(\overline{w}),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) > f_{(\boldsymbol{H}(\underline{w}),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x})$ and the lower bound on μ whenever $f_{(\boldsymbol{H}(\overline{w}),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x}) < f_{(\boldsymbol{H}(\underline{w}),\boldsymbol{X})}(\boldsymbol{h}, \boldsymbol{x})$. The lower bound for $\tau_{\overline{w},\underline{w}}$ does the reverse. □

C.6.2 Estimating bounds under linearity

Estimating the bounds without further assumptions requires nonparametric estimation of multivariate conditional means and densities. For parametric estimation, suppose the Invariant Mapping assumption and linearity hold. Then the best linear predictor, $\text{blp}(Y(w) \mid \boldsymbol{H}(w), \boldsymbol{X})$, is invariant to treatment. The following linear program then yields a plugin estimator of bounds on $\tau_{\overline{w},\underline{w}}$:

Step 1. Compute the Moore-Penrose inverses of the regressor matrices for each observed treatment level as well as $\overline{w}$ and $\underline{w}$: $\boldsymbol{Z}(w)^+ = (\boldsymbol{Z}(w)'\boldsymbol{Z}(w))^{-1}\boldsymbol{Z}'(w)$ where $\boldsymbol{Z}(w)$ is the matrix with J rows $\boldsymbol{Z}_j(w) = (\boldsymbol{H}_j(w), \boldsymbol{X}_j)$.

[54] If the intersection is empty at $(\boldsymbol{h}, \boldsymbol{x})$, mapping invariance is rejected at this point. If $\Pr(W = w \mid \boldsymbol{H}(w) = \boldsymbol{h}, \boldsymbol{X} = \boldsymbol{x}) = 1$ for any $w \in \mathbb{W}$, then $\mu(\boldsymbol{h}, \boldsymbol{x})$ is point identified (unless mapping invariance is rejected).

Step 2. Solve the linear program

$$\underline{\tau}_{\overline{w},\underline{w}} = \min_{((Y_j(w))_{j=1}^J)_{w\in\mathbb{W}}} \frac{1}{J}\sum_{j=1}^J Y_j(\overline{w}) - Y_j(\underline{w})$$
$$\text{subject to} \boldsymbol{Z}(\overline{w})^+\boldsymbol{Y}(\overline{w}) = \boldsymbol{Z}(w)^+\boldsymbol{Y}(w) \qquad \text{for } w \in \text{support}(W) \cup \{\underline{w}, \overline{w}\}$$
$$\underline{Y} \leq Y_j(w) \leq \overline{Y} \qquad \text{for } j = 1, \ldots, J,\ w \in \mathbb{W}$$

where $\boldsymbol{Y}(w)$ is the vector with jth element $Y_j(w)$, and $\underline{Y}$ and $\overline{Y}$ are bounds on potential outcomes. Also find the maximum of the same program to obtain $\overline{\tau}_{\overline{w},\underline{w}}$. Then $\underline{\tau}_{\overline{w},\underline{w}}$ and $\overline{\tau}_{\overline{w},\underline{w}}$ are estimates of lower and upper bounds on the average treatment effect.

Optionally, one can add lower and upper bounds on setting-specific treatment effects $Y_j(w) - Y_j(w')$. For instance, if $Y_j(w)$ is demand for good j at price w and one is confident that all demand curves are downward-sloping, it would be reasonable to add constraints $Y_j(w) - Y_j(w') \leq 0$ for $j = 1, \ldots, J$ and all $w \geq w'$.

C.6.3 Discretization

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

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

for all $w \in \mathbb{W}$. The average used for $\tilde{Y}_j(w)$ in the case where $W_j \neq w$ is unobserved and can range between y_{lb} and y_{ub}. 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 $W_j \neq w$, take $\tilde{Y}_j(w)$ as fixed within group. 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 the 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 3.2 in footnote 42: when discretizing the 300 observations into 60 groups, imposing $y_{\text{lb}} = 0$, $y_{\text{ub}} = 14.87$, mapping invariance, and linearity (but not no lower bound on setting-specific treatment effects), 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 a lower bound on setting-specific treatment effects, even with 120 groups the discretized problem (bounds $[1.12, 1.19]$) does not resemble the original problem.

C.7 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 $w \in \mathbb{W}$ 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.7.1 Description of the estimator

Computation of the estimator $\hat{\tau}^{\text{arb}}_{\overline{w},\underline{w}}$ 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

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

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 $w = \underline{w}, \overline{w}$. Athey, Imbens, and Wager (2018) set the tuning parameter $\zeta = 0.5$ as a default.

Step 2. Estimate the average treatment effect as

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

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 $\boldsymbol{\rho}_{\overline{w}}$ and $\boldsymbol{\rho}_{\underline{w}}$, computed in Step 1b, are meant to reduce estimation errors for $\mathbb{E}\big(\mathbb{E}(Y(\overline{w}) \mid \boldsymbol{Z}(\overline{w}))\big)$ and $\mathbb{E}\big(\mathbb{E}(Y(\underline{w}) \mid \boldsymbol{Z}(\underline{w}))\big)$ in the first term of Step 2, under the assumption of linearity.[55]

C.7.2 Theoretical result

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

Proposition 6. *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 or 3B; 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 $\hat{\tau}^{arb}_{\overline{w},\underline{w}}$ 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 $\hat{\tau}^{arb}_{\overline{w},\underline{w}}$ is asymptotically normal with*

$$\frac{\hat{\tau}^{arb}_{\overline{w},\underline{w}} - \tau_{\overline{w},\underline{w}}}{\sqrt{\hat{V}^{arb}}} \to \mathcal{N}\Big(0,1\Big)$$

where $\hat{V}^{arb} = \sum_{j=1}^{N} (\rho_{\overline{w},j} - \rho_{\underline{w},j})^2 (Y_j - \boldsymbol{Z}_j \hat{\boldsymbol{\delta}}_{\text{lasso}})^2$.[56]

Proof. The result follows from Corollary 2 of Athey, Imbens, and Wager (2018) by noting that our unconfoundedness Assumptions 3A and 3B each have the same implication as their Assumption 1 for this estimation step, our Assumptions 1, 2, and either of 3A or 3B, jointly imply their Assumption 2, and our overlap Assumption 4 is identical to their Assumption 6

[55] 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 covariates.

[56] In contrast to the variance in Proposition 1, the variance estimator $\hat{V}^{\text{arb}}$ in Proposition 6 is conditional on hypothetical evaluations. Specifically, for a fixed sample size, the weights $(\rho_{\overline{w},j} - \rho_{\underline{w},j})$ are deterministic (fixed) under sampling of outcomes Y_j conditional on covariates $\boldsymbol{Z}_j$ and treatment assignment W_j.

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 $\boldsymbol{Z}_j(\underline{w})$ and $\boldsymbol{Z}_j(\overline{w})$ and an outcome Y_j for all decision problems irrespective of treatment assignment. The two weights $\boldsymbol{\rho}_{\overline{w}}$ and $\boldsymbol{\rho}_{\underline{w}}$ 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 online appendix for the mean of the control, and the difference $\boldsymbol{\rho}_{\overline{w}} - \boldsymbol{\rho}_{\underline{w}}$ takes the role of their γ in the "Proof of Corollary 6" in their online appendix. □

C.7.3 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 $\hat{\tau}^{\text{arb},r^*}_{\overline{w},\underline{w}} = \left(\overline{\boldsymbol{Z}^{r^*}(\overline{w})} - \overline{\boldsymbol{Z}^{r^*}(\underline{w})}\right)\hat{\boldsymbol{\delta}}^{r^*} + \sum_{j=1}^J(\rho_{\overline{w},j} - \rho_{\underline{w},j})\left(Y_j - \boldsymbol{Z}_j^{r^*}\hat{\boldsymbol{\delta}}^{r^*}\right)$

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 $\hat{\tau}^{\text{arb},r^*}_{\overline{w},\underline{w}}$ 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:

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

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.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.[57]

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

[57] 10% state they have made one loan using the website, and a little over 3% state they have made two or more loans.

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, $r_{kj} \leq r^*$.[58] 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.7.3. By varying r^*, the analyst can check whether poor quality evaluations drive the results.

We illustrate this strategy in the microfinance application by filtering respondents based on correlations between their "quintile projections" and actual fundraising velocities for loan profiles displayed in their actual treatment states.[59] Figure A3 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 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

[58] 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.

[59] 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.

Table A10: 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 A3. Standard errors are conditional on the selected threshold.

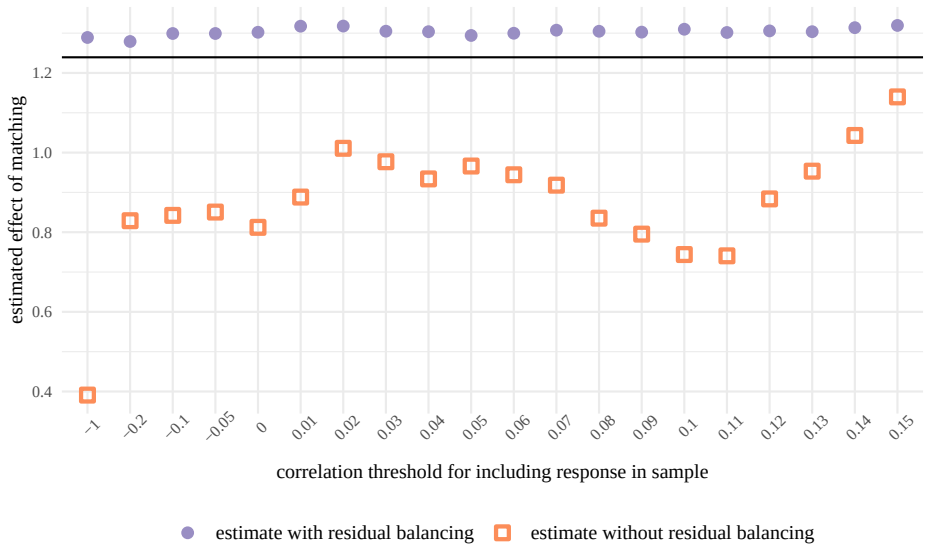
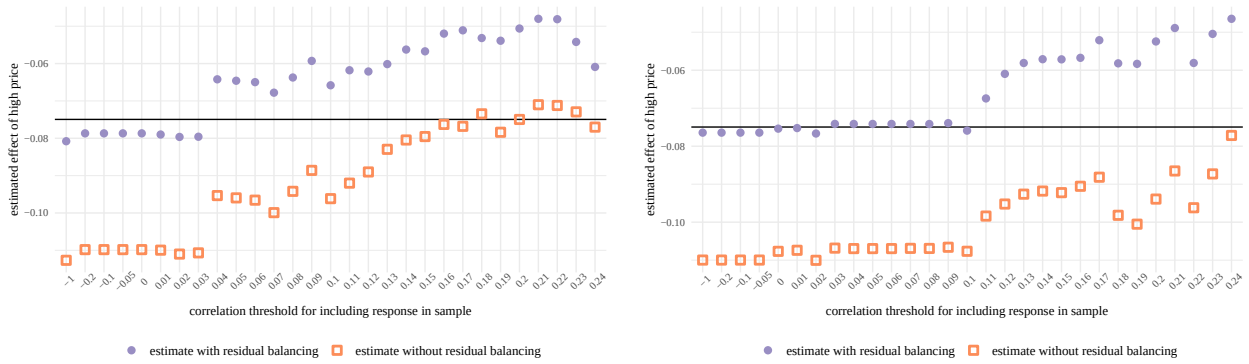


Figure A3: Estimates of the Effect of Matching by Correlation Threshold in Microfinance Application.
The horizontal line indicates the experimental estimate of the treatment effect. See Table A10 for estimates and standard errors.

Appendix C.7.3 for a detailed description of the algorithm. As the purple dots in Figure A3 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 A4. In this application, the estimates with residual balancing are mostly between -0.08 and -0.05 depending on the selected threshold, while the ground truth treatment effect is -0.075. In contrast to the microfinance application, however, the most restrictive thresholds are not closest to ground truth. For the most restrictive thresholds, the estimates without residual balancing are closer to ground truth. When the difference across thresholds between estimates with residual balancing is economically meaningful, we urge caution in choosing the threshold, as it may be challenging to determine which is optimal.



(a) Observing all snacks at high price

(b) Observing all snacks at low price

Figure A4: Estimates of the Effect of High Price by Correlation Threshold in Snack Demand Application
The horizontal line indicates the true in-sample treatment effect.

D Additional Connections to Related Literature

D.1 Relation to other methods that predict choice

Just as calibration studies in the CVM literature estimate functions relating real choices to hypothetical responses, portions of the neuroeconomics literature estimate functions relating real choices to 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.[60] Separately, in the literature on subjective well-being, several papers explore the relationships between forward-looking statements concerning happiness and/or satisfaction and hypothetical choices (Benjamin et al. 2012; Benjamin et al. 2014; Bernheim, Kim, and Taubinsky 2024), 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

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

improving those data inputs. There are also parallels to our work in the political science literature, particularly concerning the prediction of voter turnout and election results, e.g., from surveys and polls (as in Jackman 1999, and Katz and Katz 2010). As in our approach, the object is to predict aggregate outcomes rather than individuals' choices, and a range of potential predictors (in addition to hypothetical choices or intentions) are sometimes considered. For example, Rothschild and Wolfers (2011) find that questions concerning likely electoral outcomes (i.e., how others will vote) are better predictors than stated intentions.[61] The problem is substantively different, however, in that surveys and polls ask voters about real decisions that many have made, plan to make, or are in the process of making, instead of measuring non-choice reactions to choice problems that respondents view as hypothetical.

D.2 Relation to other econometric methods

D.2.1 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 qth hypothetical question.[62] 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 qth "donor unit" in period j. Further, suppose the treatment is binary, $\mathbb{W} = \{0, 1\}$. 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

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

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

synthetic controls often restrict the coefficients to be positive and to sum to 1. Synthetic control then uses the relationship to predict counterfactual 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.

D.2.2 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. 2024). 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.

E Microfinance Application

E.1 Survey Details

E.1.1 Survey Instructions

Survey participants navigated through several pages.

Page 1:

- To see if you are eligible to participate, please enter your Mechanical Turk Worker ID into the box below and then click NEXT.
- Please see below for where you can find your Worker ID. Your Worker ID starts with the letter A and has 12-14 letters or numbers. It must be all CAPITAL letters and no

spaces. It is NOT your email address.
text box

- **Please note that you cannot be paid if you enter your Worker ID incorrectly.**
- To ensure that there are no typos, please enter your Worker ID again in the box below.
text box

Page 2:

- **Protocol Director:** Professor B. Douglas Bernheim
- **Protocol Number:** IRB-42264 IRB Approval Date: October 17, 2017
- **Expiration Date:** December 31, 2099
- **DESCRIPTION:** You are invited to participate in a research study on decision-making. You will be asked to read several pages of instructions. Then you will be asked to make several choices that will determine the precise amount you will be paid, and then possibly answer several survey questions. RISKS AND BENEFITS: We cannot and do not guarantee or promise that you will receive any benefits from this study. There are no risks associated with this study.
- **TIME INVOLVEMENT:** Your participation in this experiment will take approximately as long as is indicated in the advertisement.
- **PAYMENTS:** You will be compensated at the advertised rate.
- **SUBJECT'S RIGHTS:** If you have read this form and have decided to participate in this project, please understand your participation is voluntary and you have the right to discontinue participation at any time without penalty or loss of benefits to which you are othervise entitled. You have the right to refuse to answer particular questions. Your individual privacy will be maintained in all published and written data resulting from the study.
- **CONTACT INFORMATION:**
Questions, Concerns, or Complaints: If you have any questions, concerns or complaints about this research study, its procedures, risks and benefits, you should ask the Protocol Director, Douglas Bernheim, (650) 725-8732.
Independent Contact: If you are not satisfied with how this study is being conducted,

or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Stanford Institutional Review Board (IRB) to speak to someone independent of the research team at (650)-723-2480 or toll free at 1-866-680-2906. You can also write to the Stanford IRB, Stanford University, Stanford, CA 94305-5401.

- **By continuing with this study, you are consenting to participate.**
  Please make a copy of this consent form for your own records. You can do so by right-clicking and selecting "print" in most browsers. If you cannot do so on your browser, please contact the protocol director for a copy of the consent form.

Page 3:

- On a scale from 1 to 5, how important is alleviating poverty in the developing world?
  multiple choice, single answer:
  - 1 - not at all important
  - 2 - important
  - 3 - somewhat important
  - 4 - important
  - 5 - very important
- Your answer to this the preceding question does not affect the remainder of the survey. We are interested in your opinion!
  Also, please give us your best estimates of the following. Again, your answers will not affect the rest of the survey.
- What is the population of the world?
  text box
- How many people live in poverty in the world?
  text box

Page 4:

- **Overview**: Please read all the instructions provided for this study carefully. Sometimes, you will be asked understanding questions; you must answer the first set of understanding questions correctly to complete the study.

Page 5:

- We are working with a microcredit nonprofit which alleviates poverty by providing small loans to people in developing countries.
- In this survey, we will show you profiles for actual people who have asked to borrow money. Visitors to the nonprofit's website could decide to lend small amounts of money to these individuals.
- The survey has three steps:
- First, we will show you an overview of many profiles together.
- Second, we will show you 30 profiles one by one, and ask you questions about how you would react to them as a lender.
- Third, we will show you these profiles together and ask you to make a hypothetical lending decision.

Page 6:

- **Understanding Question**: Which of the following statements is true?
  multiple choice, single answer:
  - In this survey I will play the role of a borrower
  - In this survey I will play the role of a lender
- **Understanding Question**: Which of the following statements is true?
  multiple choice, single answer:
  - This survey has two steps
  - This survey has three steps
  - This survey has four steps

Page 7 if either of the two questions on page 6 were answered incorrectly:

- We thank you for your time spent taking this survey.
- Your response has been recorded.
- *END OF SURVEY*

Page 7 if both of the two questions on page 6 were answered correctly:

- You have successfully completed the understanding questions.
- We will begin by showing you loan postings you will be asked about, before you answer any questions about them.

Page 8:

- For now, just look at the 3 postings in the top row. We'll have a closer look at the other postings later and ask you about your reactions to them.
- *Grid of loan postings. 17 rows, 3 columns. Each posting as in Figure 2, either matched or unmatched. The first row of 3 postings is identical for all participants, with the first two postings displayed without matching indicator and the third displayed with the matching indicator. The final 16 rows contain the 30 postings the participant will be asked questions about and 18 additional postings (displayed without matching indicator), in random order.*

Page 9:

- **Example Posting 1**
- The top of each posting shows you a photo of the borrower, the name of the borrower (often just their first name), and the country of the borrower. As an example, here's the first posting from the previous page.
- Overall, about 35% of loan requests on this microlending platform come from Asia, 30% from Africa, and 20% from Central and South America. The remaining loan requests come from other parts of the world, such as Oceania or Eastern Europe.
- *First loan posting from the grid, with a red box drawn around image, name, and country.*

Page 10:

- **Example Posting 2**
- In the middle of the posting, there is a brief description of what the borrower will use the loan for. In the second posting, you see that the purpose of the loan is to buy a fishing net and boat.

- Overall, about 30% of loans are for farming and agriculture, 20% for food production and sales, 15% for retail of other goods, and 14% for housing improvements and expenses.
- Although the posting shows the total amount that the borrower wants to get, most lenders each contribute $25. The contributions of many lenders add up to the total amount.
- *Second loan posting from the grid, with a red box drawn around the descriptive text.*

Page 11:

- **Example Posting 3**
- If you see the icon shown below at the bottom of the posting, that means any loan you make to the borrower is eligible for matching funds provided by one of the website's sponsors. Many of the sponsors are corporations, and some are individuals. The following explanation appears on the website:
- "...if you lend $25 to that borrower, a fund set up by the loan matcher will also lend $25 to that borrower – essentially doubling the contribution made to that loan and helping it reach its goal."
- Overall, approximately 30% of loans are matched (on the first day they are shown on the website).
- *Third loan posting from the grid, with a red box drawn around the matching indicator.*

Page 12:

- **Example Posting 3**
- Later on, we will ask you some questions about postings like this one. Please try answering the questions below.
- *Third loan posting from the grid.*
- Do you think this posting, appearing exactly as shown, attracted funds from many people? Relative to all loan postings, the number of people lending funds to this posting (per hour, during the first day while it was still accumulating funds) was: *multiple choice, single answer:*

Do you think this posting, appearing exactly as shown, attracted funds from many people? Relative to all loan postings, the number of people lending funds to this posting (per hour, during the fi rst day while it was still accumulating funds) was:

  - Low (below the 20th percentile)
  - Medium-low (20th to 40th percentile)
  - Medium (40th to 60th percentile)
  - Medium-high (60th to 80th percentile)
  - High (80th percentile or above)

- Note: If lending is matched for this posting, count each original contribution and its match as a single person.

- Receive a bonus! We will select one of the postings you see at random, and you will receive a bonus of $2 if your answer matches the number of people who lent funds per hour, during the first day while the profile was still accumulating funds.

- Imagine you regularly use a lending platform like this. How likely would you be to lend $25 to this posting?
  multiple choice, single answer:

  - 1 - very unlikely
  - 2
  - 3
  - 4 - moderately likely
  - 5
  - 6
  - 7 - very likely

- How likely would the typical user of this platform be to lend $25 to this posting?
  multiple choice, single answer, same choices as for the previous question.

- Which activity/sector would this loan fund?
  dropdown box, 16 options, single answer:

  - Agriculture, Arts, Clothing, Construction, Education, Entertainment, Food, Health, Housing, Manufacturing, Personal Use, Retail, Services, Transportation, Wholesale, Other

If an incorrect sector was selected for this posting, the following text appeared and participants had to change to the correct sector before continuing with the survey:

- This profile is actually classified as Agriculture, Food, or Retail. You can get hints by looking at the posting.

Page 13:

- Remember, we will select one of the postings you see at random, and you will receive a bonus of $2 if your answer matches the number of people who lent funds per hour, during the first day while the posting was still accumulating funds.
- *Third loan posting from the grid.*
- Your answer: *answer selected by the participant, for example,* Medium (40th to 60th percentile)
- If we had selected this posting for the bonus payment and the number of people lending to it was indeed Medium, you would get the $2 bonus.
- If we had selected this posting for the bonus payment and the number of people lending to it was Low, Medium-low, Medium-high, or High, you would not get the $2 bonus.

Page 14:

- **Understanding Question**: You will be asked about the likelihood of lending to this borrower
  multiple choice, single answer:
  - assuming you knew the borrower personally
  - based only on the borrower's posting, exactly as it appears
  - based on a transcript of an interview with the borrower
- **Understanding Question**: What does the following icon mean on a loan profile?
  image of the matching indicator.
  multiple choice, single answer:

  - this borrower is also a [website] supporter
  - if you make a loan to this borrower, you will be matched with another lender to become pen pals
  - if you decide to lend $25 to the borrower, one of the website's sponsors will match by also lending $25 to that borrower

- **Understanding Question**: If a user wants to know the country of the borrower... *multiple choice, single answer:*
  - they can "unlock" this information by purchasing a premium membership
  - they can only guess based on the name of the borrower
  - they can find it below the photo

- *If all three questions were answered correctly, advance immediately to page 18.*

Page 15 only if the first question was answered incorrectly:

- You answered the first question incorrectly. The correct statement is:
- You will be asked about the likelihood of lending to this borrower based only on the borrower's posting, exactly as it appears.
- Please answer the questions in the remainder of this survey more carefully.
- *The button to continue the survey appears only after a five second delay.*

Page 16 only if the second question was answered incorrectly:

- You answered the second question incorrectly.
- *image of the matching indicator*
  means that if you decide to lend $25 to the borrower, one of the website's sponsors will match by also lending $25 to that borrower.
- Please answer the questions in the remainder of this survey more carefully.
- *The button to continue the survey appears only after a five second delay.*

Page 17 only if the third question was answered incorrectly:

- You answered the third question incorrectly. The correct statement is:

- If a user wants to know the country of the borrower, they can find it below the photo in the overview and next to the photo in the detailed profile.
- Please answer the questions in the remainder of this survey more carefully.
- *The button to continue the survey appears only after a five second delay.*

Page 18:

- Here are the postings again. We'll ask you about 30 of them.
- Remember that you can earn a bonus by guessing which profiles attracted funds from many or few people.
- Scroll through them to get a sense for the types of profiles you think will attract lots of lenders, and the types you think will attract relatively few.
- *Grid of loan postings. 16 rows, 3 columns. Identical to page 8 except that the first row of page is skipped.*

Page 19 (30 times):

- Posting X of 30
- For the questions below, imagine that the posting appeared exactly as shown, and account for all details you consider relevant.
- *Posting X of 30*
- *Questions as on Page 12*
- *The button to continue to the next posting appears only after a five second delay.*

Page 20:

- **Results for the bonus payment**
- We randomly selected the following posting to determine whether you will receive the $2 bonus.
- *Posting selected from among those displayed in the same matching state as they appeared on the website.*
- Your answer: *participants answer, for example,* Medium-high (60th to 80th percentile)

- The correct answer: *correct answer, for example,* Medium-high (60th to 80th percentile)
- *If the answer was correct:*
  You receive the bonus. MTurk should send you a message once the bonus is on its way.
- *If the answer was incorrect:*
  You do not receive the bonus.

Page 21:

- 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?
- *Grid of loan postings. 10 rows, 3 columns. Only displaying the 30 postings reviewed in detail by the participant. Each posting has a checkbox next to it. The participant must check exactly 10 postings to continue the survey.*
- *The button to continue to the next posting appears only after a five second delay.*

Page 22:

- Finally, please answer the following questions about yourself.
- Please enter your age.
  text box
- Please select your gender.
  multiple choice, single answer:
  - Male
  - Female
  - Other
  - Prefer not to say
- Please select the level of education that you've completed.
  multiple choice, single answer:
  - Have not completed high school
  - High school diploma

  - Some college
  - College degree
  - Some graduate school
  - Graduate degree

- Please select the state in which you currently reside.
  dropdown box, 52 options, single answer: each of the 50 U.S. states; Washington, D.C.; Not in a U.S. state

- Please indicate how much you agree with the following statements.

- I made each decision in this study carefully.
  multiple choice, single answer:

  - 1 - Strongly agree
  - 2 - Agree
  - 3 - Neither agree nor disagree
  - 4 - Disagree
  - 5 - Strongly disagree

- I made decisions in this study randomly
  multiple choice, single answer, same choices as above.

- I understood what my decisions meant
  multiple choice, single answer, same choices as above.

- Have you ever visited the website *url of microfinance platform*?
  multiple choice, single answer:

  - Yes
  - No

Page 23:

- *If participant has ever visited the website url of microfinance platform:*
  Have you ever made a loan through *url of microfinance platform*?
  multiple choice, single answer:

  - No
  - Yes: 1 loan
  - Yes: 2 loans
  - Yes: 3 or more loans

- In the past 12 months, how many hours have you spent volunteering?
  text box

- In the past 12 months, approximately how much have you donated to charity? (Answer in dollars.)

- Did you draw any conclusions about our research topic? (optional)

Page 24:

- Thank you for completing the survey. Please copy the text on the following line:

- *a unique participant identifier*

- Paste it into the text box in the HIT.

- *END OF SURVEY*

E.1.2 Deviations from Pre-Analysis Plan

This experiment was preregistered (AEARCTR-0004885). Our final analysis differed from the pre-analysis plan in the following ways:

- We do not assess how well we measure treatment effects using mean-squared prediction error and calibration because the experiment only identifies average, not loan specific, treatment effects. Instead, we include p-values for tests of equality between experimental and non-experimental estimates.

- We do not estimate effects for the "secondary outcomes" listed in the pre-analysis plan. The secondary outcomes were listed as options for analysis in case estimates for the primary outcome were unexpectedly imprecise.

- We estimate the effect on compliers in the experiment using instrumental variables as is common empirical practice to avoid complications of working with transaction-level data.

- In the experiment, our funds were depleted after a total sample size (treated and untreated) of 1091, rather than the 2000 observations mentioned in the pre-analysis plan. The sample size of 500 compliers, of which 250 each treated and untreated, appears in the pre-analysis plan based on an earlier experimental design that would have randomly treated loans with even-numbered ids (odd-numbered ids untreated), rather than only those with ids ending in zero.

- We do not estimate heterogeneous treatment effects by loan sector, amount, and risk rating because the final sample size for such subgroups in the experimental sample appears too small for reliable inference. Instead, we estimate heterogeneous treatment effects for the untreated (compliers) and for the treated (always-takers). The effect among compliers is the baseline effect identified by the full experimental data set. The effect among the treated is based on our method, not the experimental data, so it is not affected by the smaller-than-expected experimental sample size.

- The pre-analysis plan says we will ask participants on Amazon Mechanical Turk to provide survey responses to questions and gives examples of 12 possibilities. After piloting, we posed fewer hypothetical questions than those listed to simplify the survey and reduce attrition. Specifically, we focused on the questions that serve as composite measures of motivations that may be predictive of the outcome, namely the prediction of fundraising velocity and a fund allocation exercise. We additionally asked respondents to rate their own and others' likelihood of lending to a given posting.

- The sample size of loans for the survey was not included in the pre-analysis plan. After some simulations for power calculations, we drew a random sample of 200 loans that were unmatched in the observational data and 100 loans that were matched. We decided to oversample matched loans to more precisely estimate the ATT using our method. Given a sample size of 300 loans, we decided to obtain about 40 evaluations for each loan under each of the two treatment states as $40 \gg \sqrt{300} \approx 17$ (a relevant comparison given our analysis of measurement error). Hence, we aimed for a total of about $300 \cdot 2 \cdot 40 = 24000$ evaluations. In the finalized survey, each respondent evaluated 30 loan postings. Hence, about $24000/30 = 800$ Mechanical Turk respondents were needed, rather than 500 as mentioned in the pre-analysis plan. Because of randomness in which loans were shown to a given respondent, we ended the survey after obtaining complete evaluations by 833 respondents.

E.2 Validation

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

E.3 Univariate models of hypothetical evaluations

E.3.1 Assuming no hypothetical bias

The four hypothetical choice variables are not directly comparable to the real choice outcome (velocity). Here, we assess whether simple strategies to convert these variables to the outcome scale yield estimates close to ground truth. Consider the two evaluations eliciting own and vicarious lending likelihood on a 7 point scale. We can convert each respondent's likelihood into a dichotomous "lending decision" by coding all responses above some threshold as "lending" and responses below the threshold as "not lending." Counting the two highest values on the 7 point scale as "lending," the average hypothetical own lending frequency among unmatched loans is 7 percentage points higher with matching than without matching. In percentage terms, lending frequencies are 43% higher on average among unmatched loans with matching than without matching. Using the vicarious lending

Table A11: Microfinance: Treatment Effect Estimates Using Each Hypothetical Response Separately

	Models using a single hypothetical variable							
	(1)	(2)	(3)	(3)	(4)	(5)	(6)	(7)
Estimated effect of matching	0.29	1.08	0.08	0.31	0.25	0.74	-0.02	0.02
	(0.26)	(0.30)	(0.14)	(0.15)	(0.21)	(0.25)	(0.16	(0.20)
projected quintile	X	X						
projected quintile, sq.		X						
own lending likelihood			X	X				
own lending likelihood, sq.				X				
other lending likelihood					X	X		
other lending likelihood, sq.						X		
included in top 10							X	X
included in top 10, sq.								X
Sample size (outcome)	300	300	300	300	300	300	300	300

Estimates of the effect of matching on the inverse hyperbolic sine of fundraising velocity, within the first day. Analytical standard errors in parentheses.

likelihood, these numbers are 9 percentage points and 41%. These estimates are much smaller than the 245% increase in funding per day implied by the experimental estimate.

E.3.2 Two step procedure with univariate specifications

Appendix Table A11 shows estimates applying our 2-step method to each hypothetical evaluation variable separately. Estimates are substantially below ground truth when using evaluations linearly. Adding second order terms yields some estimates closer to ground truth. However, without knowledge of the ground truth estimate, we would have no way of knowing which of these estimates is accurate. Models including third order or third and fourth order terms (not shown in the table) are generally similar.

F Snack Demand Application

F.1 Experiment Details

F.1.1 Groups

For the sample of snacks, we collected different responses from different groups to minimize contamination. See Appendix F.1.3 for the experimental instructions given to each group. The responses fell into three categories:

Real choices

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. This group produces the data forming our main outcome of interest.

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,[63] 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), as a robustness check.

Hypothetical choices

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.

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

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.

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

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

Group HWTP (28 subjects): Subjects expressed a hypothetical willingness to pay (WTP) for all of the food items, each of which appeared only once. We employed this protocol because much of the literature explores the accuracy of hypothetical WTPs rather than binary choices. We used the same subjects for groups HWTP and L (below).[64]

Subjective responses

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 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

[64]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).

approval"). These ratings are intended to capture social norms and image concerns.

Group S (29-38 subjects):[65] Subjects answered some or all of the following additional questions concerning the food items (answers scaled 1-5): 1) "How much would you later regret eating this snack?" 2) "How tempting is this item?" 3) "If you had no concerns about diet or health, how much would you enjoy eating this item?" 4) "Is this item generally good or generally bad for you?" 5) "Would others form a positive or negative impression of you if they saw you eating this snack?" 6) "Are people likely to understate or overstate their inclination to pick this snack?" The responses to these questions may be useful for predicting choices because each question potentially measures factors related to the degree of hypothetical bias. Questions 1 through 4 address the degree to which immediate gratification conflicts with longer term considerations: we conjectured that hypothetical choices will be more sensitive to long-term costs, and less sensitive to immediate gratification, than real choices. Question 5 addresses concerns for social image: we conjectured that hypothetical choices will be more sensitive to image concerns than real choices. Finally, question 6 may determine whether subjects can provide subjective assessments of hypothetical bias that would be useful for the purpose of predicting choices, even if the sources of the bias remain unclear.

F.1.2 Session-level aggregates included in specifications with detailed hypothetical evaluations

Specifications with detailed hypothetical evaluations as regressors (such as Columns (5) and (6) of Table 3) include, first, a set of price-specific variables aggregated from the sessions described in the previous section:

- 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)

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

- the 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 F.1.1; 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 either 45 or 46 base session-level variables.

F.1.3 Experimental Instructions

The instructions given to each group took a similar form, with small adjustments.

Hypothetical group The materials for the hypothetical group (H) follow. Session timing (T=T1+T2) varied by the task assigned to each group, and show up fee (F1) and payment (F2) were occasionally altered depending on demand to attract a sufficient number of respondents.

Script Participants are assigned to a seat as they enter the room. As they are seated, tell each the following:

Please read and fill out the consent form; the second copy is for you to keep. Then, read the instructions, making sure you answer the questions at the end. If you have any questions about either form, let me know. Please don't start on the computer until we instruct you to.

Get their first names as you seat them, and check them off the participant list.

If more than 10 people arrive on time, seat the first 10. Give the remainder the $F1 show-up fee and ask them to come back another time.

If one participant is over 5 minutes late, begin without them. Don't allow late arrivals to participate. If they are late, they do not have a claim to any payment.

Once all participants are assembled, introduce the project:

Hello and welcome. Today you will get the chance to participate in a study on food choices. Your participation will help us understand how people make this type of basic decision.

If at any point during the experiment you decide you do not wish to participate, please let us know. The activities will take approximately T minutes. As a thank you for your participation, you will receive $F2.

If you complete the first part before everyone else is finished, you may work on a quiet activity at your desk. We will let you know when it is time to move on to the second part.

Before we start, does anyone have any questions?

Participants may decline to participate at any time; they should not be forced to do anything.

If a participant declines to participate for some reason, they will receive their full payment.

If a participant asks a question that is not part of the main script, but does not impede his/her ability to solve the task, tell them, "I'm sorry, but I cannot answer that right now, since we must carefully control the information we give out."

To start the experiment:

• Write down the subject number and session ID on their instruction sheet

• Enter the same numbers in the prompt on the computer

• Ensure that the instructions come up, and then tell them to read the instructions before continuing

Part 2:

• Give all subjects the questionnaire. Distribute the snacks as applicable.

• At 4:30, go to the subjects one-by-one. Check that they have filled out all the forms, and have them collect their payment.

Sheet The experiment has two parts.

- During the first part, you will make choices on a computer (this will take approximately T1 minutes).
- During the second part, you will answer some questions on paper (this will take approximately T2 minutes.

At the end of the experiment you will receive $F2 as compensation for your participation in the experiment.

Instructions for Part 1

In each trial you will be shown a picture of a food snack on a computer screen, and you will have to indicate whether or not you would be willing to pay either $0.25 or $0.75 to be able to eat one serving of that snack during the second part of the experiment. Keep in mind that these choices will be hypothetical. That is, you will not actually receive any of the items. However, your answers are still very important for the success of our experiment, so we thank you for taking the questions seriously.

You should press the 1 key for "yes" and the 2 key for "no". To avoid mistakes, please keep your fingers on the keyboard at all times.

Treat each food independently; that is, think of each food as the only one you might receive, regardless of your previous answers.

If you are not sure what constitutes a serving size for that food, please raise your hand to ask the experimenter.

You will have as much time as you need to make your decision, but you will not be able to advance to the next item until the current one has been on your screen for 4 seconds. That means there is no incentive to rush through the questions. Once you have finished, you may work on a quiet activity until the next part.

Instructions for Part 2

During this part you will be asked to fill out a short written questionnaire. Your answers to these questions will not affect in any way the payment you receive at the end of the experiment.

If you finish early, you will be free to work on homework or any other quiet activity, but you must stay in the room until you are dismissed.

Do you have questions about the instructions or activities? If so, please ask the experimenter. If not, please answer the following question:

• Multiple choice: Which possible prices will you see during Part 1?

Circle up to 2:

• $0.25
• $0.50
• $0.75
• $1.00

Onscreen Welcome to the study.

You will see pictures of many food items, one at a time.

For each food item, answer the following hypothetical question.

Would you pay $x (where $x will either be $0.25 or $0.75) to receive one serving of this item as a snack during part 2 of the experiment?

If yes, press "1". If no, press "2".

Treat each food independently; that is, think of each food as if it is the only one you might receive, regardless of your previous answers.

Remember, the question is hypothetical -- you will not actually buy this item.

We have placed $2 in quarters next to this computer. It is part of your $X payment for participating.

If you are not sure what constitutes one serving size, please ask the study leader.

You will have as much time as you like to answer the question, but the slide will not advance until at least 4 seconds have passed.

To reduce errors, it is recommended that you keep your fingers on the keyboard and look at the screen.

Press any key to continue.

Adaptations by group We report differences in the materials between groups in this section. Meaningful changes or additions to the materials for group H are designated in italics.

Group R

Script "you will receive *an envelope containing about $F2, with the exact amount depending on the decisions that you make*."

"*Transition to part 2:*

Once everyone has finished the computer part, say the following:

> *Everyone has finished the first part of the experiment. Before we move on to the second part, we need to determine which snack you will get, if any. I will flip a coin to determine which half of the slides we will be selecting from. Heads is the first half; tails is the second half. You will then draw a number between 1 and 200 to determine which slide is selected. If you said yes to the food on that slide, you will*

get that food in the second half. If you get a food, I will point out that food on your screen. I will also take the appropriate payment from your quarters.

Determine the foods:

- *Flip the coin in view of two of the subjects.*
- *Have each subject draw from the randomization box and show you the number.*
- *Find the trial corresponding to that number in My Documents\Choice Experiments\group1_hl*.txt.*
- *If they said no (RESP=2), move on.*
- *If they said yes (RESP=1), highlight that food and point it out to them, and take the appropriate number of quarters. Write down the item they should receive.*

Once you have finished this part, tell all subjects the following:

Please gather all of your belongings, as well as your forms, pens, and quarters. Follow us to the first floor for the second part."

Sheet

- "During the second part, you will answer some questions on paper (this will take approximately 20 minutes)*, and depending on your choices in part 1, you may also receive a snack that you will be allowed to eat during this time.*"
- "At the end of the experiment you will receive *either $F2, $F2-0.25, or $F2-0.75 (depending on your choices in the first part)* as compensation for your participation in the experiment."
- "In each trial you will be shown a picture of a food snack on a computer screen, and you will have to indicate whether or not you *are* willing to pay either $0.25 or $0.75 to be able to eat one serving of that snack during the second part of the experiment."
- Deleted: *"Keep in mind that these choices will be hypothetical. That is, you will not actually receive any of the items. However, your answers are still very important for the success of our experiment, so we thank you for taking the questions seriously.*"
- *"At the end of the experiment we will select one of the trials at random and will implement your decision. Thus, if we randomly select a trial in which you say yes, you will receive the snack shown in that trial, and the appropriate charge will be deducted from your $X participation fee."*

- You will have *a minimum of 4 seconds* to make your decision. That means there is no incentive to rush through the questions. Once you have finished, *please wait until an experimenter comes to assist you. Important: Since you don't know in advance which trial will be selected, you should treat every decision as if it were the only one."*

- *"After part 1 is completed, you will follow the experimenter to another area, where you will be put in separate rooms."*

- "You will have to stay in the room for 20 minutes for this part...During this time you may eat any food purchased as part of the experiment, but you may not eat anything else. Important: If you have a medical condition that might require you to eat something during that 20 minute period, it is extremely important for you to notify a staff member before the experiment begins."

- "...please answer the following *questions: ● If I am hungry during the experiment and did not choose the snack that was randomly selected, I may eat a snack I brought with me or buy one from a vending machine. Circle one: True False ● If I only answer "Yes" to one snack item, it is more likely that that item will be drawn. Circle one: True False*"

Onscreen

- *"At the end of the experiment we will select one trial at random and implement your decision for that trial. Since you don't know in advance which trial will be selected, you should treat every decision as if it were the only one."*

- "We have placed $2 in quarters next to this computer. It is part of your $X payment for participating. *If you end up buying an item, you will pay for it out of this money.*"

<u>Group M</u>

Script *"Determine the foods:*
Have each subject draw from 1-5 from the randomization box and show you the number.
Find the trial corresponding to that number in txt file.
If they said no (RESP=2), move on.
If they said yes (RESP=1), highlight that food and point it out to them, and take the appropriate number of quarters. Write down the item they should receive.
Part 2:

Once everyone has finished the computer part, say the following:

Everyone has finished the first part of the experiment. We will now pass out the questionnaire for part 2.

If you get a food, I will bring it to you know. I will also take the appropriate payment from your quarters."

Sheet

- "During the second part, you will answer some questions on paper (this will take approximately 20 minutes)*, and depending on your choices in part 1, you may also receive a snack that you will be allowed to eat during this time."*

- *"For all but five of the food items*, you will have to indicate whether or not you would be willing to pay either $0.25 or $0.75 to be able to eat one serving of that snack during the second part of the experiment. Keep in mind that these choices will be hypothetical. That is, you will not actually receive any of the items. However, your answers are still very important for the success of our study, so we thank you for taking the questions seriously."

- *"Please note that FIVE (and only five) of your decisions will actually be REAL choices. These will be clearly marked, so that you know which ones they are. At the end of the experiment we will select one of these REAL choices at random and implement that decision. Important: Since you don't know in advance which of the five REAL choices will be selected, you should treat every decision as if it were the only one."*

- *During this time you may eat any food purchased as part of the study, but you may not eat anything else. Important: If you have a medical condition that might require you to eat something during that 20 minute period, it is extremely important for you to notify a staff member before the experiment begins.*

- "...please answer the following *questions: 1. If I only answer "Yes" to one REAL choice, it is more likely that that item will be drawn. Circle one: True False 2. Fill in the blank: I will see _____ REAL choices; the rest will be HYPOTHETICAL.*"

Onscreen

- "For *all but five of the* food *items, you will* answer the following hypothetical question."

- Added: "*--- page break ----Please note that FIVE (and only five) of your decisions will actually be REAL choices. These will be clearly marked, so that you know which ones they are. At the end of the experiment we will select one of these REAL choices at random and implement that decision. Since you don't know in advance which of the five REAL choices will be selected, you should treat each of them as if it were the only REAL choice. We have placed $2 in quarters next to this computer. They are part of your payment. If you end up buying an item, you will pay for it out of this money.*"

Group HCT

Script Prior to participants making choices, the experimenter read the additional statement:

> *Before you begin the experiment, we would like to make you aware of a problem we often have in studies like this one. As we have explained, you will be asked to make hypothetical choices, not real ones. You will not actually pay money in exchange for a snack at the end of the experiment. But we have also asked you to respond to the choices as though you are really deciding whether to purchase the snack.*
>
> *And that creates a problem.*
>
> *In most studies of this kind, people seem to have a hard time thinking about hypothetical choices. They choose differently when the choices are hypothetical, and they don't really have to pay money to buy something, than they do when the choices are real, and they really can buy something.*
>
> *For example, in a recent study, a number of subjects made the same types of hypothetical choices you will be making here. No one actually had to pay money even if they said they would be willing to buy an item at a particular price. Another set of subjects made real choices involving the same items. They really did have the chance to buy the same items at the same prices. The result was that the percentage of items people said they would buy in response to hypothetical questions was 15 percentage points higher than the percentage of items other people actually bought when they faced real choices. That's quite a difference, isn't it?*
>
> *We call this a "hypothetical bias." "Hypothetical bias" is the difference that we usually see in the way people respond to hypothetical choices compared to real choices.*

How can we get people to think about hypothetical choices like they think about real choices, where if they say they're willing to buy something, they actually buy it? How do we get them to think about what it means to really dig into their pocket and pay money, if in fact they really aren't going to have to pay anything?

Here's why we think we see this hypothetical bias, why people make hypothetical choices differently than real choices. When we are asked whether we would buy something like a tasty snack in a hypothetical situation, we think to ourselves, sure, I like that item, so yes, I would buy it. We may even think hard about whether or not we want a particular item at a particular price, and say to ourselves "yes, I really would buy that item now, I really think I would," but still answer based on whether we like the item, rather than on whether we would actually pay money for it. If that's the case, then answers to hypothetical questions only tell us whether people like the items that are offered, and "liking something" is not the same as "buying something" when a purchase decision is real.

When the choice is real, and we actually have to spend our money to buy the item, we think differently. We may still think the item is desirable, but we think about whether we really want or need it right now, and whether we really want to spend our money that way. For example, if we buy a snack now, that's money we don't have to buy a snack or a soft drink later today, or a cup of coffee later at a coffee shop. So when the choice is real, we choose in a way that takes into account the limited amount of money we have -- even when the item we are considering doesn't cost very much, like a snack or a soft drink. This is just a theory, of course, but it's what we think may be going on with hypothetical choices.

In any case, the only way that we know of to get people like you to make hypothetical choices the same way you would make real choice is simply to ask you the following. In the hypothetical choices that you'll start making in a few minutes, please think hard about what you're choosing. Ask yourself, if the choice were real – if you really were deciding whether to buy that item (and only that item) for the stated price – would you actually want to spend the money to buy it, right here and now? If you would, then indicate yes, and if you wouldn't, indicate no. In short, we ask you to choose exactly as you would if you were really going to face the consequence of your choice, which is to either pay the price and have the snack during the second half of the experiment, or not pay the price, keep your money for something else, and have no snack during the experiment.

Please keep this in mind throughout the experiment.

Sheet Unchanged.

Onscreen Unchanged.

Group HL Note that when we use the responses of this group we recode the choices so that higher values are more likely and that responses range between 0 and 1.

Script Unchanged.

Sheet

- "In each trial you will be shown a picture of a food snack on a computer screen, and you will have to indicate *how likely* you would be to pay either $0.25 or $0.75 to be able to eat one serving of that snack during the second part of the experiment."
- "You should *answer on a scale from 1 to 5, with 1 = very likely, 3 = uncertain, and 5 = very unlikely.*"

Onscreen

- "*How likely is it that you would* pay $x (where $x will either be $0.25 or $0.75) to receive one serving of this item as a snack during part 2 of the experiment?"
- "*Answer on a scale of 1 to 5, with 1 = very likely, 3 = uncertain, and 5 = very unlikely.*"

Group HV Note that when we use the responses of this group we recode the choices so that higher values are more likely and that responses range between 0 and 1.

Script. Unchanged.

Sheet

- "Instructions for Part 1 *This part consists of 2 rounds. In each round, you will be asked a question about a food item. You will then see many food items, presented one at a time. Please answer the question for each food item using the number keys. Important: We*

thank you in advance for rating your perceptions accurately, as this will be a great help for our research. After you have finished answering the questions, we ask that you sit quietly at your computer until time is up. Rushing through the questions will not affect how long the experiment lasts. Once it is time for Part 2, the study leader will come and assist you."

- "please answer the following *question: Fill in the blank: In Part 1, there are __ rounds."*

Onscreen *"This part of the study consists of 2 rounds. In each round,* you will see pictures of many food items, one at a time. For each item, answer the question *you are asked on the screen. To enter your response, use the number keys.*" Members performed the same task as Group HL, except they were asked to indicate how they thought a typical undergraduate of their own gender would answer.

Group HWTP/L Note that when we use the responses of group L we recode the choices so that higher values are more likely and that responses range between 0 and 1.

Script Added, "*Finally, please read the instructions on the computer carefully. Note especially that for the first round you will need to enter your responses in cents.*"

Sheet "Instructions for Part 1 *You will see many food items, presented one at a time. Please answer the question you are asked for each food item using the number keys. There will be two rounds, with a different question in each round. Please note: In the first round, you will be asked to enter a price in cents. So, if your desired response is $0.25, you should type 25 and then press the enter key. Important: We thank you in advance for rating your perceptions accurately, as this will be a great help for our research. After you have finished answering the questions, we ask that you sit quietly at your computer until time is up. Rushing through the questions will not affect how long the experiment lasts. You are free to work on homework, however. Once it is time for Part 2, an experimenter will come and assist you.*"

On screen *"The first part of the study consists of 2 rounds.*

In the first round, you will be shown a series of pictures of snack items. In each case, imagine that you will be offered an opportunity to buy this item to eat as a snack during the second part of this experiment, and that the alternative is to have no snack. Hypothetically, what is the maximum amount you would be willing to pay for it?

For example, if you answer that 10 cents is the most you are willing to pay for an item, that means you would want to buy the item and eat it as your only snack during the second part of the experiment if the price turned out to be less than 10 cents (say 9 cents), but that you would prefer to have no snack during the second half of the experiment if the price turned out to be more than 10 cents (say 11 cents). Likewise, if you answer that $1.50 is the most you are willing to pay for an item, that means you would want to buy the item and eat it as your only snack during the second part of the experiment if the price turned out to be less than $1.50 (say $1.42), but that you would prefer to have no snack during the second half of the experiment if the price turned out to be more than $1.50 (say $1.63).

Remember that these are hypothetical choices, and that they will not be implemented. However, we encourage you to answer as honestly as possible.

You will enter your answer in cents. So, if your answer is $1.25, you will enter 125. Press the enter key to submit your response.

The directions for the second round will be given later.

To reduce errors, it is recommended that you keep your fingers on the keyboard and look at the screen.

Press any key to continue.

--- insert WTP questions ---

In this round, you will see pictures of many food items, one at a time.

Think about receiving each item you are shown as a snack during the second part of the experiment.

Then answer the following question:

How much would you like to eat this item during the second part of the experiment, on a scale of 1 to 7, with 1 being not at all, and 7 being very much?

To enter your response, use the number keys."

Group SWB

Sheet "Instructions for Part 1

You will see many food items, presented one at a time. Please answer the question you are asked for each food item using the number keys.

Important:

- *We thank you in advance for rating your perceptions accurately, as this will be a great help for our research.*

- *After you have finished answering the questions, we ask that you sit quietly at your computer until time is up. Rushing through the questions will not affect how long the experiment lasts. You are free to work on homework, however. Once it is time for Part 2, an experimenter will come and assist you.*"

- Deleted: multiple choice questions.

Onscreen *"You will be shown a series of pictures of snack items, each one appearing along with a price.*

In each case, imagine that you receive this item (and ONLY this item) to eat as a snack during the second part of this experiment, and that the indicated price will be deducted from your show-up payment. Answer the following question:

How happy would be on with that outcome on a scale of 1 to 7, where 1 indicates very unhappy, and 7 indicates very happy?"

Group N

Script Added to the end of part 1: "*Determine the foods After each subject finishes, open their data file and see if they said yes or no to the snack. Write down responses on the roster.*"

Sheet "Instructions for Part 1

You will see many food items, presented one at a time. Please answer the question you are asked for each food item using the number keys.

Important:

- *We thank you in advance for rating your perceptions accurately, as this will be a great help for our research.*

- *After you have finished answering the questions, we ask that you sit quietly at your computer until time is up. Rushing through the questions will not affect how long the experiment lasts. You are free to work on homework, however. Once it is time for Part 2, an experimenter will come and assist you.*"

- Deleted: multiple choice questions.

Onscreen *"In each case, we want you to imagine that a subject in this experiment paid a specified price to eat the item as a snack during the second part of the experiment. We would like you to indicate whether you think the typical person would approve or disapprove of this purchase.*

In particular, answer the following question:

Imagine that a subject in this experiment paid XX 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?

Please answer on a scale of 1 to 7, where 1 indicates strong disapproval and 7 indicates strong approval."

Group S *Different sessions asking these questions used slightly different variants, as shown in brackets (see footnote 65).*

Script Unchanged.

Sheet

- "Instructions for Part 1 *This part consists of {6,3,2} rounds. In each round, you will be asked a question about a food item. You will then see many food items, presented one at a time. Please answer the question for each food item using the number keys. Important: We thank you in advance for rating your perceptions accurately, as this will be a great help for our research. After you have finished answering the questions, we ask that you sit quietly at your computer until time is up. Rushing through the questions will not affect how long the experiment lasts. Once it is time for Part 2, an experimenter will come and assist you."*
- "please answer the following *question: Fill in the blank: In Part 1, there are __ rounds."*

Onscreen *"The experiment consists of {6,3,2} rounds. In each round,* you will see pictures of many food items, one at a time. *Think about choosing each item you are shown as an afternoon snack. Then answer the question you are asked on the screen. To enter your response, use the number keys."*

F.2 Assessing whether respondents take the "real choice" seriously

In the "mixed" group (M), 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 (R), 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.

F.3 Quantifying "hypothetical noise"

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

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

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

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

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

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

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

Rearranging, we have

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

To bound σ_ω^2, note that by the law of total variance $\sigma_\omega^2 = \mathrm{var}(\omega_j) = \mathrm{var}(\mathbb{E}(\omega_j \mid \mu_j)) + \mathbb{E}(\mathrm{var}(\omega_j \mid \mu_j))$. 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 $\mathrm{var}(\omega_j \mid \mu_j) = \mu_j(1-\mu_j)/N$. 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 (2014) found that the variance of valuations rises when choices become less consequential.

F.4 Overlap assessment

For all applications, we recommend producing overlap plots such as those shown in Figure A5 to assess the potential stability of equation 2. 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. 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).

F.5 Univariate models of hypothetical evaluations

F.5.1 Assuming no hypothetical bias

Columns (2) through (6) of Table A4, 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, but rely on the conventional assumption that hypothetical responses reveal the choices people would actually make, at least up to a constant ($\mu_w^{conventional}(h) = h + b$ for some b; see Section 2.2.1).

As noted in Section 4.1.3, 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 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.

F.5.2 Two step procedure with univariate specifications

In panel (b) of Table A4, 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 the simulation assumes real choices are observed at the high price or 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 not surprising due to the failure of overlap.

F.6 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. Specifically, we set

$$W_{js} = 1\left\{\mathrm{WTP}_j > \epsilon_{js}\right\},$$

for item j in simulation s, where random shocks ϵ_{js} are independent draws from a t-distribution.[66] 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.[67] Because the analyst typically would not have access to those surveys, we do not include hypothetical WTPs in the vector $\boldsymbol{H}_j$. Appendix Figure A6 shows there is predictive relationship between each snack's actual purchase frequencies (potential outcomes) and the simulated probability it is observed at

[66] 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.

[67] 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 A7 is -0.078 instead of -0.077.

the high price. Alternative assignment mechanisms yield qualitatively similar conclusions.

Table A7 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 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 3. 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.7.

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.

Correcting for sample selection changes the median estimates using low-dimensional specifications (Columns (4) and (5)) to -0.087 (both specifications; from -0.081 and -0.077, respectively). Bounds on the average treatment effect without assuming mapping invariance yield a range $[-0.543, 0.457]$. Imposing mapping invariance and linearity according to the specification in Column (4) yields a range $[-0.231, 0.226]$. For the specification in Column (5), the range is $[-0.259, 0.226]$. Additionally imposing that demand is (weakly) downward-sloping for each snack yields ranges $[-0.158, -0.035]$ and $[-0.163, -0.030]$, respectively. While there is some variation of these bounds across samples, the true average treatment effect of -0.075 is contained in the sets for all samples.

F.7 Estimation of heterogeneous effects

Even in settings with random treatment assignment, our method may reveal heterogeneity in treatment effects at a level of granularity that is difficult to achieve using real choices alone. In this section, we explore this possibility by comparing 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, $\tau_j = Y_j(1) - Y_j(0)$:

- *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.[68]

Results. In this section, we begin with simulated environments where treatment is assigned randomly (with the high price prevailing for 94 of 189 snacks) and ones where there is no variation in treatment.

[68] 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."

Results appear in Figure A7. For each estimation method, we plot each metric's median value and interquartile range based on 10,001 simulated samples.

We begin with two benchmarks. The first—a simple difference-in-means estimator (row 1)—does not try to capture heterogeneity but is unbiased for the average effect so long as treatment is random.[69] Because the 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 estimator may perform well relative to alternatives in terms of MSE and simulated profits by virtue of its parsimony.

The second benchmark measures the greatest amount of heterogeneity one might hope to capture by conditioning on the physical characteristics linearly. Specifically, row 2 reports performance for the best linear predictor of actual setting-level treatment effects based on all the physical characteristics. This approach would be infeasible in applications because it requires observations of each setting in both treatment states.[70] Because it predicts treatment effects as linear functions of physical characteristics and nothing else, feasible estimators involving functions of hypothetical variables could in principle outperform it, even when similarly restricting attention to linear predictors and using either no real data or a single real observation per setting.

In rows 3, 4, and 5, we consider three common 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. These common approaches are applicable when treatment assignment is random, but not when treatment variation is absent.

Alternately, a conventional 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 estimator yields substantially higher R^2 and allows simulated firms to achieve higher profits than the common estimators based on physical characteristics. These results highlight the fact that hypothetical evaluations contain much

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

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

Table A12: Summary Statistics Describing Heterogeneous Treatment Effect Estimates in Snack Demand Application

	Benchmarks		Observational			Our Method				
	Diff. in Means	True TE Proj. on Char.	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 Means" does not capture any heterogeneity) and for the column using projected true effects ("True TE Proj. on Char.") 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.

useful information. However, hypothetical choice bias leads to high mean squared error.[71]

Our method (rows 7-10) combines hypothetical evaluations and observations on choices, which allows it to capture much heterogeneity while removing the hypothetical choice bias, thereby reducing mean squared error and further improving calibration and profit. These rows of Figure A7 show results for the variant that employs hypothetical choices and physical characteristics as predictors (i.e., the same variant as in Table 3 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 as described in Appendix F.6.

Our method performs substantially better across the board than the three feasible common estimators that do not use hypothetical evaluations. It also easily surpasses the infeasible best linear predictor 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 this application, even though unadjusted hypothetical choices do not reflect that variation accurately. Because our method does not require an intervention, it can enable analysts to recover heterogeneous treatment effects even when they lack the power to intervene. This feature may be particularly valuable in settings where one wishes to target the treatment at those who are likely to benefit most.

F.8 Gains in precision

We consider two standard estimators: difference-in-means and ARB (with physical characteristics entered linearly along with 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 3.

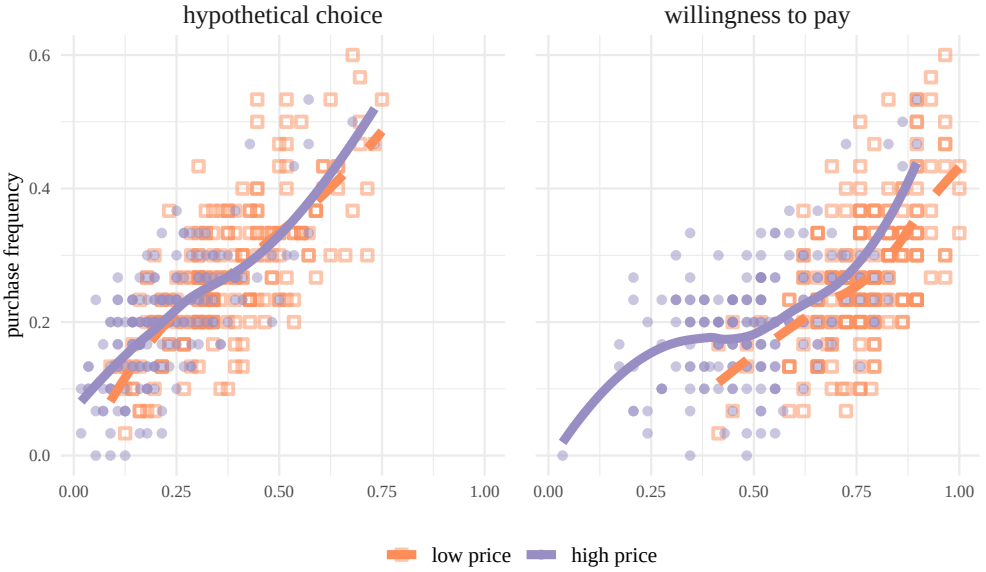
The standard deviations of our estimators are substantially smaller than those of the standard estimators, especially when the proportion treated is far from half, as shown in the first panel of Figure A9. The standard deviation of standard estimators is U-shaped in the

[71]The other hypothetical variables mostly perform worse except with respect to mean squared error, as shown in Appendix Figure A8.

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. 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, standard 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, and is more precise. 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 setting.

In this application, a smaller standard deviation comes at the cost of a small bias (Figure A9 second panel), but our estimators attain lower root-mean-squared error, irrespective of the treatment's prevalence (Figure A9 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.

(a) Overlap in Hypotheticals



(b) Relationship between Outcome and Hypothetical

Figure A5: Overlap between Hypothetical Evaluations in Snack Demand Application

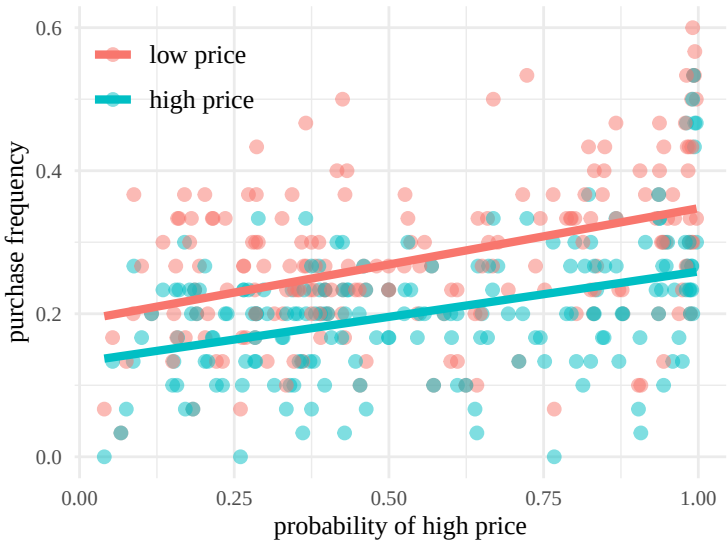


Figure A6: Endogenous Snack Pricing Simulation Setup

Potential outcomes corresponding to the high price are in blue, and potential outcomes corresponding to the low price are in red. 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.

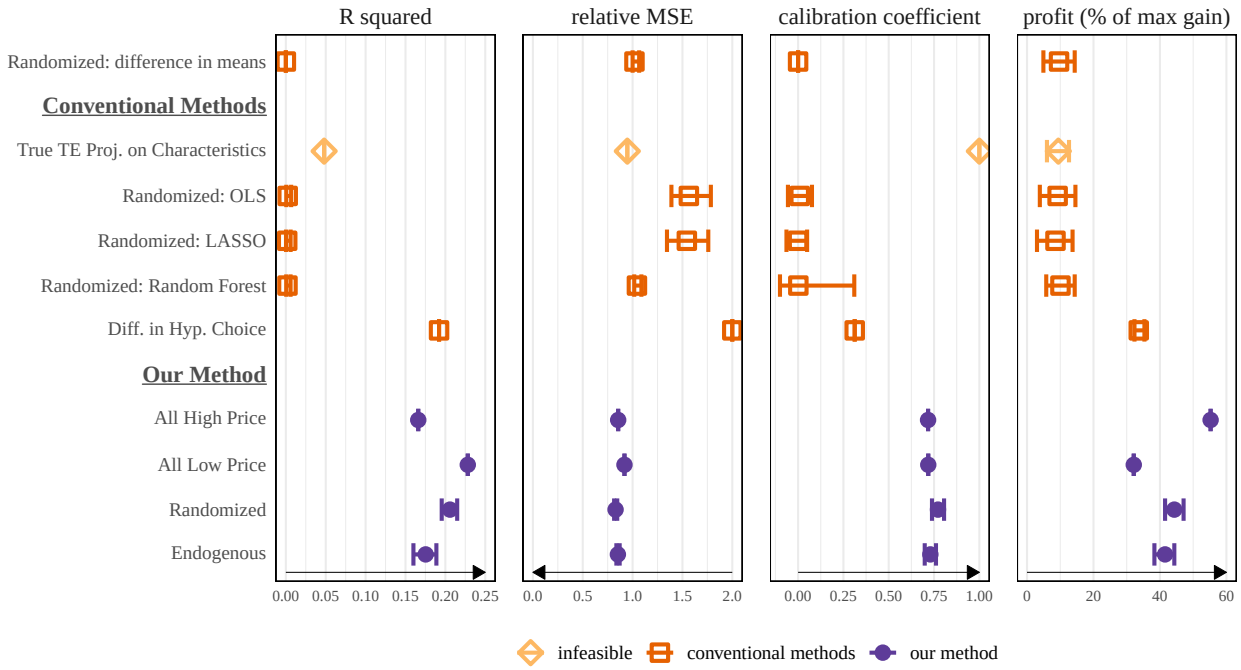


Figure A7: Treatment Effect Heterogeneity in Snack Demand Application

Summary statistics describing how well different estimators capture heterogeneity in treatment effects, per R^2 and three additional measures. Mean squared error (MSE) is computed relative to predicting with the ATE $\frac{\text{mean}((\tau_j-\hat{\tau}_j)^2)}{\text{mean}((\tau_j-\tau)^2)}$. Calibration coefficient (the slope coefficient in a regression of τ_j on $\hat{\tau}_j$) measures whether the expectation of the actual treatment effect increases unit for unit with the predicted treatment effect; the ideal coefficient is unity. For 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 F.7. Points indicate the median value of each statistic across 10,001 simulated samples, and whiskers indicate the interquartile range. For the relative MSE, the value for "Diff. in Hyp. Choice" is 3.62 (by construction not varying with treatment assignment) but is shown in the figure as 2 because the axis is truncated. 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. Row 1 presents a difference in means estimate, which does not try to capture heterogeneity but is unbiased for the average effect so long as treatment is random. Row 2 reports performance for the best linear predictor of actual setting-level treatment effects based on all the physical characteristics: this is infeasible but represents the greatest amount of heterogeneity one might hope to capture by conditioning on the physical characteristics linearly. Rows 3, 4, and 5 consider common 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. Row 6 represents a conventional hypothetical approach, which yields decently high R^2 but hypothetical choice bias leads to high MSE. Rows 7-10 show that our method achieves good performance on all metrics, regardless of how treatment is assigned. See Appendix Table A12 for additional specifications.

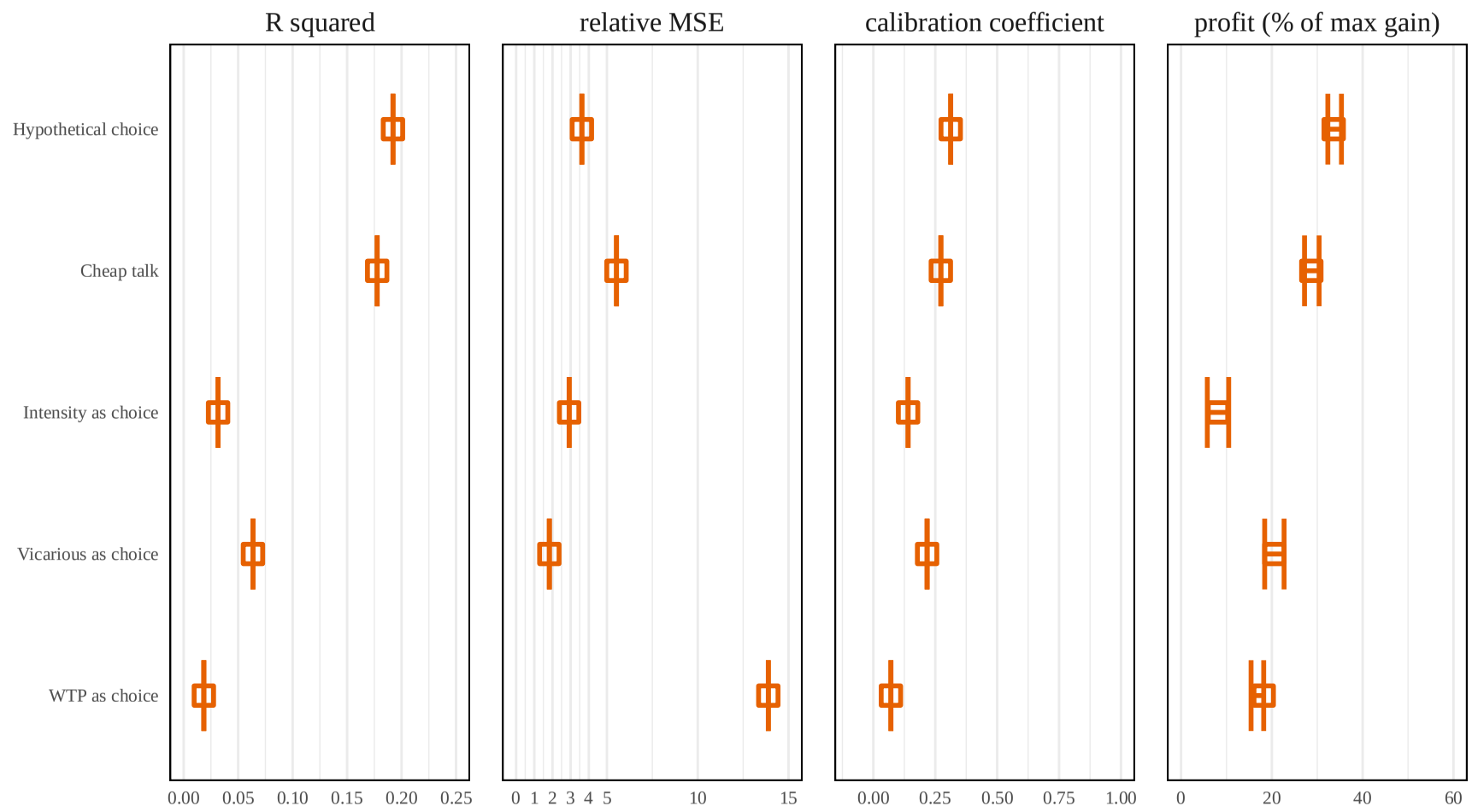


Figure A8: Treatment Effect Heterogeneity Using Solely Hypotheticals As Predictions in Snack Demand Application

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 4.1.2. Points indicate the median value of each statistic across 10,001 simulated samples, and whiskers indicate the interquartile range.

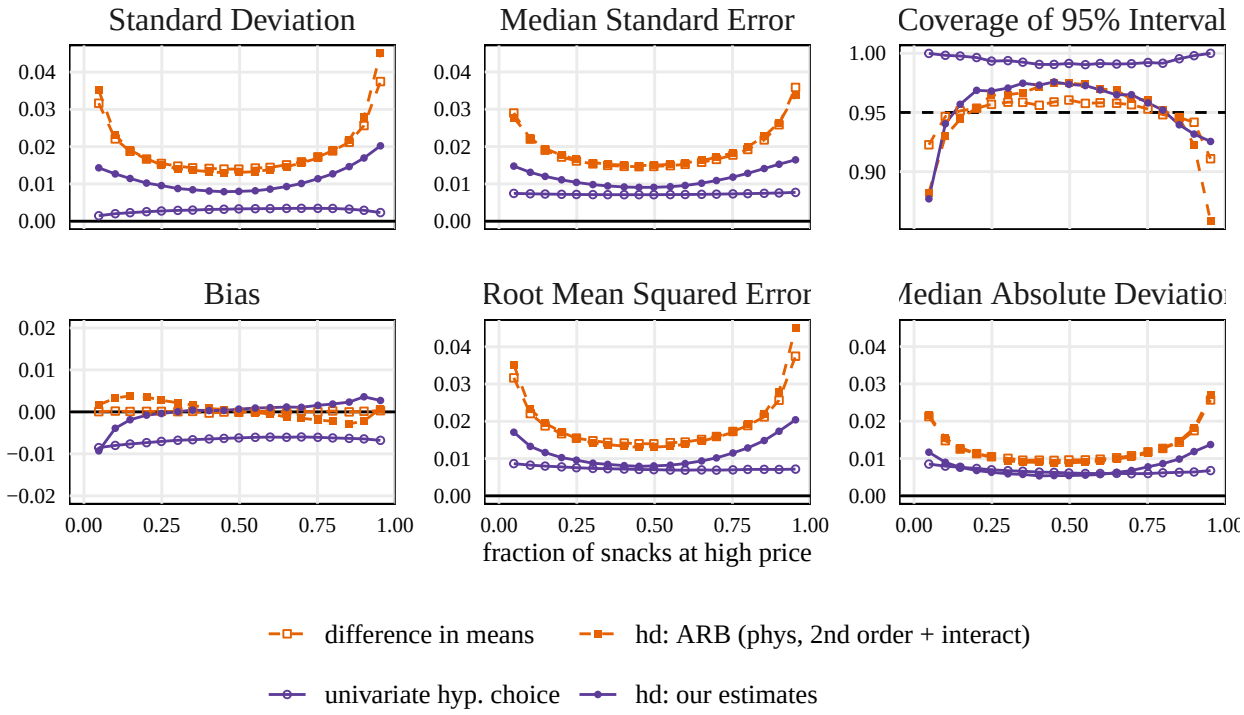


Figure A9: Performance of Estimators by Fraction Treated in Snack Demand Application

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 4.2.1), and the standard deviation across assignment realizations is mechanically zero because there is only one possible assignment. Coverage of confidence intervals refers to coverage of the true in-sample treatment effect. For additional specifications see Appendix Figures A10 and A11.

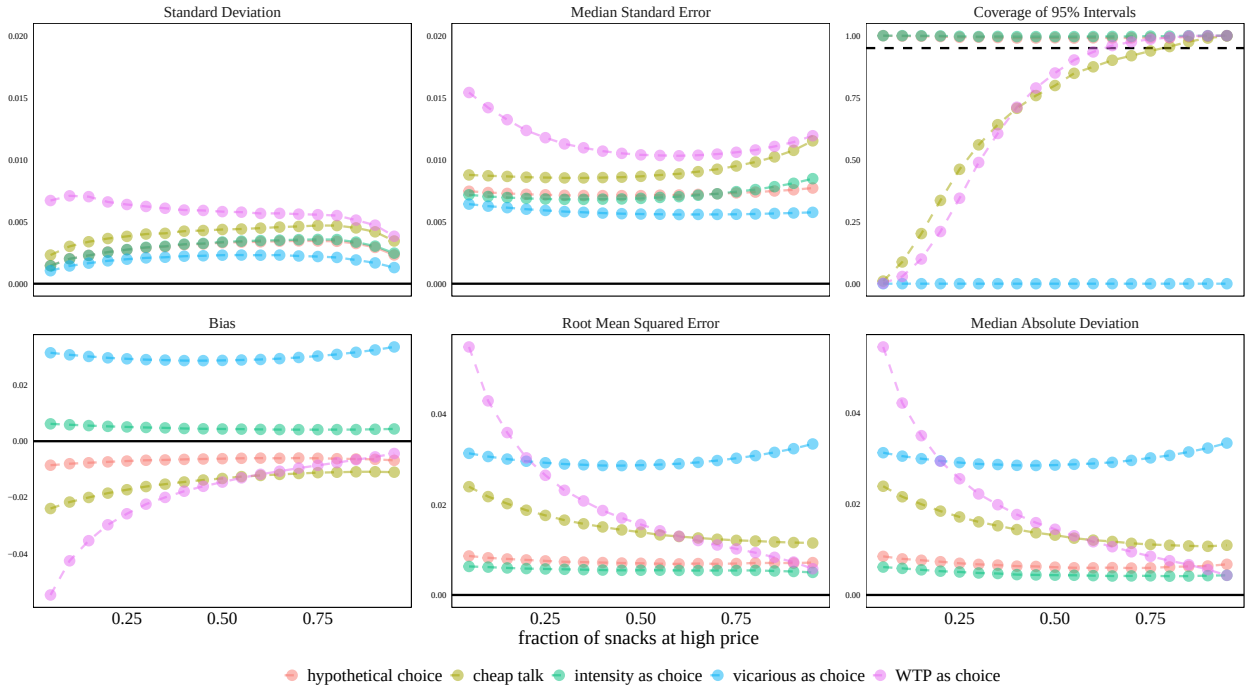


Figure A10: Performance of Univariate Estimators by Fraction Treated in Snack Demand Application

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 Appendix Table A4b. Coverage of confidence intervals refers to coverage of the true in-sample treatment effect.

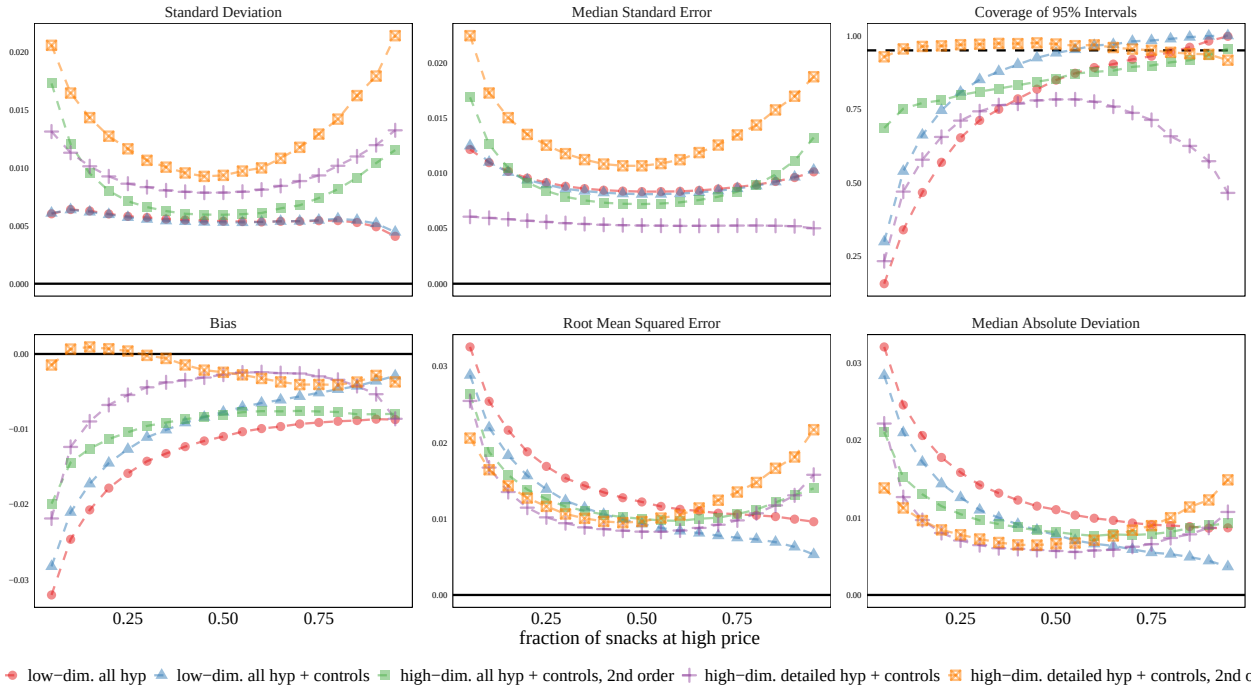


Figure A11: Performance of Multivariate Estimators by Fraction Treated in Snack Demand Application

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 3 in the main text. Coverage of confidence intervals refers to coverage of the true in-sample treatment effect.