Manipulation-Robust Prediction: Online Appendix

Daniel Björkegren Joshua E. Blumenstock Samsun Knight

August 5, 2026

Contents

S1 Estimation Extensions **3**
S1.1 Moment Conditions for Greenfield Case . . . 3
S1.2 Moment Conditions for Brownfield Case . . . 3
S1.3 One-Shot Estimation . . . 4

S2 Experimental design **10**
S2.1 Intake . . . 10
S2.2 Study treatments and interventions . . . 10
S2.3 Communicating decision rules . . . 12
S2.4 Attrition management . . . 12
S2.5 Timeline . . . 13
S2.6 Selection of outcomes and behaviors . . . 14
S2.7 Deviations from pre-analysis plan . . . 21

S3 Can surveys measure manipulation costs? **27**
S3.1 Intake . . . 28
S3.2 Survey design . . . 28
S3.3 Results . . . 29
S3.4 Validation . . . 30

S4 Supplemental results **31**
S4.1 Testing model assumptions . . . 31
S4.2 Additional comparative statics . . . 33

S4.3 Additional monte carlo simulations 34
S4.4 Additional performance comparisons 34
S4.5 Anticipating fit 35

S1 Estimation Extensions

We list moment conditions formulated in the case where shocks have been decomposed, $\boldsymbol{\epsilon}_{it} = \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}$, and common shocks $\boldsymbol{\mu}_t$ are absorbed with time fixed effects. They also apply without time fixed effects, in which case one may omit $\boldsymbol{\mu}_t$ and swap $\boldsymbol{\epsilon}_{it}$ for $\boldsymbol{\eta}_{it}$. These systems are overidentified, and we weight moments equally.

S1.1 Moment Conditions for Greenfield Case

Implemented decision rules are orthogonal to idiosyncratic behavior shocks and manipulation noise ($\mathbb{E}[\beta_{itk}\eta_{itj}] = 0$ and $\mathbb{E}[\beta_{itk}v_i] = 0$). For each pair of behaviors jk (including $j = k$) this yields sample moment condition

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\mathbb{T}_i}\beta_{itk}\left[x_{ijt} - \underline{x}_{ij} - \mu_{jt} - f(\boldsymbol{\omega}, \mathbf{z}_i)\cdot[C^{-1}\boldsymbol{\beta}_{it}]_j\right] = 0 \tag{1}$$

where $[\mathbf{a}]_k$ indicates the kth element of $\mathbf{a}$.

One can form an estimate of unobserved heterogeneity $\hat{v}_i$ by

$$\hat{v}_i = \frac{1}{\sum_{t\in\mathbb{T}_i^{treatment}}|K_{it}^{eval}|}\sum_{t\in\mathbb{T}_i^{treatment}}\sum_{k\in K_{it}^{eval}}\left[\frac{x_{ikt} - \underline{x}_{ik} - \mu_{kt}}{[C^{-1}\boldsymbol{\beta}_{it}]_k} - f(\boldsymbol{\omega}, \mathbf{z}_i)\right], \tag{2}$$

where K_{it}^{eval} is the set of behaviors to be evaluated for i in period t.[1] Unobserved heterogeneity is mean zero, yielding moment condition, $\frac{1}{N}\sum_i \hat{v}_i = 0$, and orthogonal to each heterogeneity characteristic, yielding moment condition(s) $\frac{1}{N}\sum_i z_{li}\cdot\hat{v}_i = 0$ for each characteristic l.

S1.2 Moment Conditions for Brownfield Case

In greenfield settings where the base decision rule $\boldsymbol{\beta}_0 = \mathbf{0}$, it is possible to infer natural behavior from baseline behavior (e.g., equation (5)), prior to estimating costs. Our method can also be applied in *brownfield* settings, where a decision rule has already been implemented and baseline behavior may already be manipulated.

In such a setting, one can jointly estimate costs and the parameters describing baseline behavior ($\underline{\mathbf{x}}$ and $\boldsymbol{\mu}$) by appending two moment conditions based on $\mathbb{E}[\eta_{itk}] = 0$. For each

[1] We set $K_{it}^{eval} = \{k \text{ s.t. } \beta_{itk} \neq 0\}$, so that $\hat{v}_i$ is evaluated on shifts in the incentivized behavior.

individual i and behavior k, we have

$$\hat{\underline{x}}_{ik} = \frac{1}{|\mathbb{T}_i|} \sum_{t \in \mathbb{T}_i} \left[x_{ikt} - \mu_{kt} - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_k \right] \tag{3}$$

For each time period t and behavior k we have

$$\hat{\mu}_{kt} = \frac{1}{|\{i|\mathbb{T}_i \ni t\}|} \sum_{i|\mathbb{T}_i \ni t} \left[x_{ikt} - \underline{x}_{ik} - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_k \right] \tag{4}$$

Identification still requires observing random variation along each behavior in the decision rule (and ensuing manipulation).[2]

S1.3 One-Shot Estimation

In our experiment, we observed each individual over multiple time periods, which increased statistical power. However, our approach can also be applied if each individual i is observed in only one period t (for example, if loan applicants each apply for a loan once).

In a one-shot setting, C and $\boldsymbol{\omega}$ can be recovered by adjusting the brownfield moment conditions to remove both individual and time fixed effects. This entails replacing the moment condition in equation (1) with

$$\frac{1}{N} \sum_{i=1}^{N} \beta_{itk} \left[x_{ijt} - \chi_j - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_j \right] = 0,$$

Equation (2) with

$$\hat{v}_i = \frac{x_{ik_it} - \chi_{k_i}}{[C^{-1}\boldsymbol{\beta}_{it}]_{k_i}} - f(\boldsymbol{\omega}, \mathbf{z}_i),$$

Equation (3) with

$$\hat{\chi}_k = \frac{1}{N} \sum_{i=1}^{N} \left[x_{ikt} - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_k \right],$$

and dropping equation (4), where natural behaviors are replaced with a term representing common behavior $\boldsymbol{\chi}$ of dimension K, and where k_i is the behavior incentivized for individual i.

An estimate of each person's natural behavior can then be obtained by undoing any

[2]This inversion will be more sensitive to the specification of the model than when unincentivized behavior can be observed directly in training.

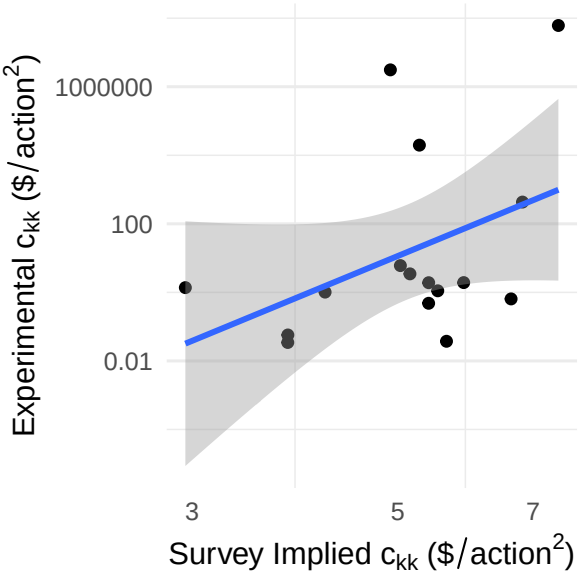
predicted manipulation:

$$\underline{\hat{x}}_{ik} = x_{ikt} - f(\hat{\boldsymbol{\omega}}, \mathbf{z}_i) \cdot [\hat{C}^{-1}\boldsymbol{\beta}_{it}]_k$$

though with just one observation this will be more affected by idiosyncratic noise.

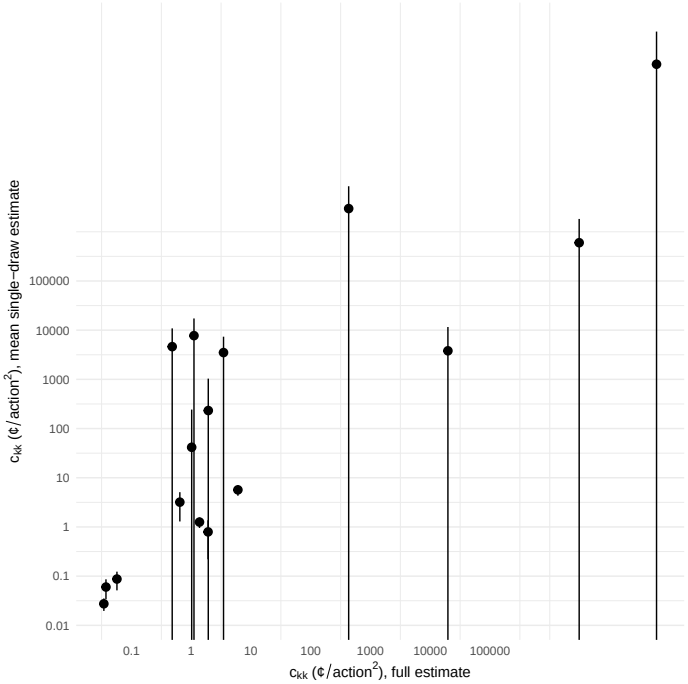
We demonstrate that this approach obtains similar results by mimicking the data that would result if our experiment had observed each individual for only one period. Because this will drastically reduce our sample size, we simulate this over multiple replication draws. For replication draw r, for each individual i we restrict the sample to include only one randomly selected incentivized week $t_{ir} \in \mathbb{T}_i^{treatment}$, and consider the average over replications $r \in \{1...R\}$. Figure S2 shows that the one-shot estimates are similar to the full sample estimates (we report the average $c_{kk} = \frac{1}{R}\sum_r c_{kkr}$); the Pearson correlation coefficient between the two measures is 0.9987. The corresponding estimate for $\omega = \frac{1}{R}\sum_r \omega_r = 0.22$ (standard deviation 0.98), with 30.7% of single-draw estimates less than or equal to the full-sample estimate of -0.083 and 69.2% of estimates greater than the full-sample estimate.

Figure S1: Costs Elicited from Hypothetical Questions and Costs Measured in Experiment



Notes: Each dot represents a behavior captured by the Sensing App. Y-axis indicates the cost of manipulating that behavior, estimated through our experiment (Table 4). X-axis indicates costs elicited from hypothetical questions, inferred as $c_{kk} = \frac{1}{N_{survey}} \sum_i \frac{\beta_k}{\max(0.001, \Delta_{kki})}$ for each i surveyed (see Online Appendix S2.3).

Figure S2: Manipulation Costs Estimated with only One Observation per Person



Notes: Our main estimates (with multiple observations per person) are shown on the x-axis. The average estimate obtained when each individual is observed only once is shown on the y-axis using the one-shot moment conditions in Appendix S1.3. The standard deviation across replications is shown as a whisker in either direction.

Table S1: Estimated Manipulation Costs for All Behaviors

Heterogeneity by Behavior (**C** diagonal; all incentivized behaviors)

▲	●	text you send	0.03
▲	●	text you receive	0.04
		text you send on a weekday	0.05
		each text message you send in the evening hours (after 6pm)	0.05
▲	●	text you send or receive in the evening (6pm−10pm)	0.06
▲	●	time you call someone	0.48
		call with a number not in your contacts during the workday (9am−5pm	0.59
	●	second of your shortest weekend call	0.64
		call with a number not in your contacts on the weekend	0.81
▲	●	each different person you text or are texted by	1.02
		person you text during the early morning hours (12am − 5am)	1.04
▲	●	time you receive a call	1.11
		each person you text in the evening (after 6pm)	1.38
		person you text on the weekday that you send the most texts	1.55
		call between 12am and 5am	1.67
		each different person you text or are texted by on the weekend	1.68
		text you send on the day that you send the most texts	1.68
		person you call during the workday	1.70
	●	call you make that's missed	1.91
▲	●	call with someone not in your contacts	1.93
		person you receive a text from during the evening (5pm − 10pm)	1.98
		text you send on the day you send the least texts	3.03
▲		text you receive on the day you receive the most texts	3.47
		if you keep it at 100%.	3.87
		person you text	5.45
		outgoing call on the day with the least outgoing calls	5.77
▲		person who texts you	6.04
		person who calls you during the workday	6.67
		each person you call for over 30 minutes	7.12
		person you text on the workday that you text the most people	7.57
		time you plug in your phone and it charges	9.39
		if the number of texts you send varies a lot weekday to weekday	28.15
		person you receive a text from during the early morning hours (12am − 5am)	35.06
		day you use a whatsapp−related app	35.44
		if the number of texts you send and receive varies a lot day to day	92.37
		missed call on the day you have the least missed calls	112.94
		tools−type app you use	194.41
		if the number of texts you receive varies a lot weekday to weekday	217.33
		5−minute−or−more call with a non−contact	426.58
		day you use the gmail app	431.59
		day you use the app	519.97
		if you use a whatsapp−related app.	594.14
		if you use a tools−type app.	1,055.35
		if you use any document− or report−related app.	1,194.08
		productivity−type app you use	1,383.08
		if you use the gmail app.	10,349.72
		if you use a productivity−type app.	17,087.82
▲	●	second of your average evening (6pm−10pm) call	19,761.46
		if you use any twitter−brand app.	20,775.27
		text you receive on the day you receive the least texts	25,618.26
		missed call	26,820.04
		text you receive on the evening (6pm−10pm) that you receive the least texts	39,063.42
		if you use the app	111,824.80
		second of your average workday call (monday−friday	346,246.83
		if you use the twitter app at least once	370,683.79
		call that lasts at least an hour	395,022.17
▲	●	second of your average call duration	3,108,632.19
		day you use any instagram−brand app	7,505,847.37
		day that you use a productivity−type app	8,462,705.30
		instagram−brand app used	28,377,875.40
		social−type app you use	61,264,134.39
		document or report−related app you use	75,097,239.06
		times the average day−to−day variation in your texts	1,282,497,178.36
		percentage point of your time that you spend within .5 kilometers of the busara office	184,130,822,787.78

Notes: Parameters estimated using GMM. In cost matrix, off diagonal elements c_{jk};j$\neq k$ regularized to zero ($\Lambda_{offdiagonal} \to \infty$), diagonal elements regularized with $\Lambda_{diagonal} = 1.0$, set via 3-fold cross validation. ▲ : included in incentivized naïve LASSO decision rule, ● : included in incentivized strategy-robust (SR) decision rule.

Table S2: SR Decision Rules Based on Survey-Estimated Costs

	Costs (Actual)	*Costs* (From Survey)	**Income** $\boldsymbol{\beta}^{LASSO_{final}}$	**Income** $\boldsymbol{\beta}^{SR_{final}}_{SurveyCost}$	**Income** $\boldsymbol{\beta}^{SR_{final}}$	**Intelligence** (above median Ravens) $\boldsymbol{\beta}^{LASSO_{final}}$	**Intelligence** (above median Ravens) $\boldsymbol{\beta}^{SR_{final}}_{SurveyCosts}$	**Intelligence** (above median Ravens) $\boldsymbol{\beta}^{SR_{final}}$
Panel A: **Decision Rule**								
text_count_out	0.035	3.804	-0.499	-0.329	-0.101			
text_count_incoming	0.037	5.645	0.141	0.014		0.270	0.223	0.115
text_count_evening	0.057	3.805			-0.120			
call_count_out	0.480	5.4	0.657	0.591	0.531		-0.058	
call_count_outgoing_missed	1.914	5.4				-0.156		
calls_noncontacts	1.929	5.891				-0.547		-0.519
max_daily_texts_incoming	3.471	5.155						0.420
Intercept			296.342	305.309	302.472	489.686	483.530	487.033
$\lambda^{decision}$			759.296	759.296	759.296	1032.37	1032.37	1032.37
Panel B: **Prediction Error**				RMSE ($)			RMSE ($)	
Predicted Opaque			3.572	3.577	3.584	4.972	4.982	4.973
Predicted Transparent			3.876	3.644	3.591	4.988	4.989	4.975

Notes: Panel A reports the decision rules derived from naive LASSO and our strategy-robust model, as well as strategy-robust decision rules that use only control weeks and costs estimated from surveys. It also reports the costs associated with these behaviors. Panel B reports the predicted performance of these decision rules, based on the experimentally estimated model of behavior. $\beta^{LASSO_{final}}$ presented in this table differs slightly from the β^{LASSO} which was implemented. The regularization protocol was updated to select penalization closer to the boundary of 3 coefficients and the sample was changed to coincide with that used for the SR model (it includes only individuals with nonmissing tech skills, dropping approximately 1.5 percent of the sample). For survey costs, we infer heterogeneity in gaming ability using variation in participant responses (see Online Appendix S2).

S2 Experimental design

S2.1 Intake

The subject population was Kenyans aged 18 or older who owned a smartphone and were able to travel to the Busara center in Nairobi. Busara assembled a list of prospective participants from prior experiments and from market centers in Nairobi. Prospective participants were called by phone and invited to participate in an intake session in the Busara lab.

At the lab, participants were briefed on the project and software, and given a chance to ask questions. Those who consented were assisted in installing the app on their phones, and instructed on the procedure of the research project: they were given a walk-through of the app, shown how to upload data, given an explanation about the behavioral challenges and monetary incentives, and assigned a mock challenge to test the system on their devices. At the end of this session, participants were asked to complete an intake survey.

For more detailed information on intake, including call scripts and step-by-step descriptions of the procedures, see `https://dan.bjorkegren.com/manipulation-appendix-extra.pdf`

S2.2 Study treatments and interventions

During each active week of the study, each participant was randomly assigned a different challenge. Participants were notified of these challenges via text message (SMS) and through a phone-based notification that directed them to open the app. When a user opened the app, the initial screen offered them an opportunity to participate in that week's challenge, but did not show them the specific challenge that they had been assigned to. If they accepted, the app showed their particular challenge for that week, examples of which are given below. Participants were paid for their participation, plus any rewards they earned through the challenges, on a weekly basis.

During **control** weeks (**C**), participants did not receive incentives to change their behavior. Instead, they received a message of the form, *Dear user, you do not have to do anything for this week's challenge. You will receive an extra Ksh 50 for accepting this challenge. This will be paid next week Wednesday with the rest of your payout if you upload data in the upload window.*

By contrast, **simple challenges** incentivized a specific behavior monitored by the app (treatment S_k, for $k \in \{1, ..., K\}$, where k indexes the features assessed). Examples include:

- *You will receive 35 Ksh for each day you use a photography app this week, up to Ksh.*

1000!

- *You will receive 3 Ksh for each text message you send in the evening hours (after 6pm) this week, up to Ksh. 1000!*
- *You will receive 10 Ksh for each call you receive during working hours (9am-5pm) this week, up to Ksh. 1000!*
- *You will receive a base payment of 100 Ksh but from that we will* ***deduct*** *3 Ksh for each call you receive during working hours (9am-5pm) this week.*

For these challenges, we considered a decision rule of the form $\pi(\mathbf{x}_{it}) = \beta_{itk}x_{itk} + \alpha_{it}$ Ksh. The incentive amount β_{itk} was drawn at random from the set $\{-2r_k, -r_k, r_k, 2r_k, 4r_k, 8r_k\}$, for a scalar r_k that was selected for each behavior so that the maximum payout was triggered by the 90th percentile of baseline behavior.[3] Thus, most individuals were paid for a particular behavior (positive incentive), but some were penalized (negative incentive). All decision rules included a constant term (a 'participation payment') α_{it}; it was 50 Ksh if there were no negative coefficients, or if there was at least one negative coefficient we set it to a higher number, usually 250 or 450 Ksh. The final payment was capped between 50 and 1000 Ksh, so that participants always got at least a participation payment (that is, the actual payment was $\max(50, \min(\pi(\mathbf{x}_{it}), 1000))$).[4]

Finally, **complex challenges** ($\mathbf{X}$) describe a particular outcome that was being predicted; for example, *This week, earn up to 1000 Ksh. if the Sensing app guesses that you are a high-income earner based on your calls, texts, and app use.* For these challenges, we considered decision rules of the form $\pi(\mathbf{x}_{it}) = \boldsymbol{\beta}_{it}\mathbf{x}_{it} + \alpha_{it}$. The final payment was capped between 50 and 1000 Ksh, so that participants always got at least a participation payment (that is, the actual payment was $\max(50, \min(\pi(\mathbf{x}_{it}), 1000))$).[5] Participants were randomly assigned to one of several outcomes, under two cross randomized conditions: either a naïve (N) or robust (R) decision rule, and either a opaque (O) or transparent (T) condition. Under the opaque condition, no additional information about the decision rule was provided. Note that under opacity, naïve and robust decision rules are observably equivalent to users until the point

[3] For a small number of challenges where the maximum payout was predicted to be quite high at the level of $8r_k$, we slightly reduced the highest payout.

[4] Prior to October 2019, the maximum payout was scaled to 250 Ksh rather than 1000 Ksh.

[5] The constant term did not affect incentives for any behaviors, so for complex challenges, after the first week we heuristically 'rounded down' the payout function's α_{it} from the estimated decision rule constant in order to conserve financial resources, assigning a rounded constant term (either 10 or multiples of 50, up to 300) that would still ensure a positive payout (in case of negative incentives) at the 90th percentile of behavior.

of payment. Thus, for each outcome-week, our analysis pools the opaque treatments that paid out according to different decision rules. Under the transparency condition, users were presented with two additional interface elements. First, users were given the mathematical formula used for their payout that week, in natural language. For example, *Your payment this week will be 200 Ksh., minus 1 Ksh. for every 10 texts you send, minus 1 Ksh. for every 10 texts you send or receive in the evening (6pm-10pm), plus 5 Ksh. for every 10 calls you make.* Second, users were provided a screen with interactive sliders that computed expected payouts from different hypothetical amounts of the relevant behaviors (see Figure 2 in the main paper).

To limit the potential for individuals to learn from previous weeks, we ensured that no participant would receive an opaque complex challenge after receiving a transparent challenge for the same outcome.

S2.3 Communicating decision rules

In focus groups, we found that people had difficulty understanding decimals or complicated mathematical operations. To make the rules more interpretable, we formatted decision rules as follows:

- Each coefficient was rounded to the nearest integer. If the nearest integer was zero, the denominator was inflated by factors of 10 until it became nonzero. For example, we sent '3 Ksh per 10 text messages sent' rather than '0.3 Ksh per text message sent.' If the unit was seconds or minutes, the denominator was instead inflated by factors of 60.[6]

- The constant term was reported last, unless the first coefficient was negative, in which case the constant was reported first.

To mitigate order effects, for challenges with multiple features, we randomized (at the individual level) the order in which features were listed.

S2.4 Attrition management

Attrition had two dimensions: some participants did not upload data through their app, and some did not opt in to the weekly challenges. Although the app continuously uploaded data in the background, our participants had spotty internet access, so we also included a

[6]When computing predicted behavior based on estimated decision rules, we did not round coefficients.

manual upload button and emphasized the importance of being connected to the internet and uploading during a weekly window (between 1pm Tuesday and 10am Wednesday).

To minimize both types of attrition, participants were sent regular reminders via text message. Participants received a text every Tuesday at 1 pm to remind them to upload their data through the Smart Sensing app. Participants who had not yet uploaded data or activated their challenge were contacted by phone and surveyed by Busara:

- On Wednesday at noon, participants who had not uploaded any data during the end-of-week upload window were contacted.
- On Thursday and (if not responsive to the first outreach) a second time on Friday, participants whose phones did not confirm with our servers that the challenge was received or opted in to were contacted and surveyed.

Any participant who did not answer a call on the first attempt was re-contacted once more by the surveyor. We restrict analysis to participant-weeks in which the participant opted in and uploaded during the end-of-week upload window.

S2.5 Timeline

The study had three phases: 'Phase 0' load-in weeks; 'Phase 1' simple challenge weeks; and 'Phase 2' complex challenge weeks.

- **Phases 0 and 1: Weeks 1 - 19**

  Phase 0 was the period when participants were on-boarded by the Busara team. Given the capacity of the lab, approximately 200 people were on-boarded each week.

  Over the course of the study, we several times used the control data collected up to that point to decide which features to assess in simple challenges (that is, to estimate the manipulation costs of). We began by attempting to predict various outcome variables (listed below) with LASSO models that selected up to 3 features. We restricted consideration to focal outcomes that could be well predicted, and chose their models' constituent features, as well as alternative features (close substitutes).[7]

  After onboarding and 1 week of control challenges, participants in Phase 1 were each week randomly assigned either control (**C**) challenges or simple (**S**) challenges. These challenges incentivized a feature that the research team determined to be predictive

[7] We define alternatives in Section S2.6.3.

of a possible Y outcome given the baseline data available at that point. We typically assigned roughly 1/3 of our onboarded sample (i.e., the set of participants who had already been assigned their initial control week challenge) into 'control' challenges every week.

In order to offset attrition, an additional on-boarding period began in week 12 and continued through week 14, adding approximately 1000 'top up' participants to the study. As with the original cohorts of participants, these were assigned control challenges during their first week and then randomly assigned either simple challenges and control challenges in subsequent weeks.

- **Phase 2: Weeks 20-27**

  Phase 2 was the period when 'complex' challenges were assigned, which targeted the Y outcomes.

  No challenges were assigned between December 19, 2019 and January 1, 2020, in the weeks covering Christmas and New Year's.

S2.6 Selection of outcomes and behaviors

We developed a set of candidate outcomes and behavioral features, then narrowed down the subset of features for which to measure manipulation costs in Phase 1 and the subset of outcomes to incentivize in Phase 2.

S2.6.1 Candidate outcomes

We considered candidate outcome variables from variables gathered in the baseline survey and from phone activity observed in participants' first control week in the study. Below, we list the unscaled $\mathbf{y}^{unscaled}$ versions of all the outcome variables we considered. When we trained decision rules, we scaled the selected variables to obtain the outcomes $\mathbf{y}$.

- Survey-based outcomes
  - Self-reported number of acquaintances (both raw and arcsinh transformed)
  - Self-reported number of close friends (both raw and arcsinh transformed)
  - Self-reported number of contacts in phone contact list (both raw and arcsinh transformed)
  - Self-reported number of Facebook friends (both raw and arcsinh transformed)
  - Self-reported number of calls made in the week prior to taking the survey (both raw and arcsinh transformed)

- Self-reported number of calls received in the week prior to taking the survey (both raw and arcsinh transformed)
- Self-reported number of people spoken to in the week prior to taking the survey (both raw and arcsinh transformed)
- Self-reported number of texts received in the week prior to taking the survey (both raw and arcsinh transformed)
- Self-reported number of texts sent in the week prior to taking the survey (both raw and arcsinh transformed)
- Self-reported number of people relying on participant (both raw and arcsinh transformed)
- Self-reported cellular data usage in the week prior to survey
- Self-reported airtime spent in the week prior to survey
- Self-reported phone sharing frequency: weekly or above
- Popularity PCA, derived from principal components analysis of self reported variables: an indicator for having a Facebook account, calls received week prior to survey, texts sent week prior to survey, texts received week prior to survey, number of people in contacts, number of people spoken to week prior to the survey, number of acquaintances, number of close friends
- Assets PCA, based on principal components analysis of self reported variables: an indicator for owning an iron, the number of habitable rooms in their house, the number of mosquito nets in their possession, the number of towels owned, the number of frying pans owned, and an indicator for having an electric socket
- Education level: primary or below (self-reported)
- Education level: some college or above (self-reported)
- Access to electric socket (self-reported)
- Gambling frequency: weekly or above (self-reported)
- Gender (self-reported)
- Year of birth above median (self-reported)
- Number of household members (self-reported)
- Number of children (self-reported)
- Any children (number of children > 0) (self-reported)
- Marital status: married (self-reported)
- PHQ-9 (Patient Health Questionnaire-9, a 9-item mental health screening survey) mental health category: minimal
- PHQ-9 (Patient Health Questionnaire-9, a 9-item depression screening survey) mental health category: mild
- PHQ-9 (Patient Health Questionnaire-9, a 9-item depression screening survey) mental health category: moderate
- PHQ-9 (Patient Health Questionnaire-9, a 9-item depression screening survey) mental health category: moderately severe
- PHQ-9 (Patient Health Questionnaire-9, a 9-item depression screening survey) mental health category: severe
- Monthly income (both raw and arcsinh transformed)

- Monthly income above zero
- Monthly income above 25th percentile
- Monthly income above 50th percentile
- Monthly income above 75th percentile
- Addition skills, measured by whether they correctly solved a word problem requiring addition to solve
- Division skills, measured by whether they correctly solved a word problem requiring division to solve
- Exponentiation skills, measured by whether they correctly solved a word problem requiring exponentiation to solve
- Ravens test (non-verbal intelligence test) score: number correct
- Ravens test (non-verbal intelligence test) score: above median
- Occupation: employed/student
- Occupation: cleaner/house help
- Occupation: driver
- Occupation: housewife
- Occupation: sales person
- Occupation: small business
- Occupation: student
- Occupation: teacher
- Occupation: unemployed and searching for a job
- Occupation: unemployed and not searching for a job
- Occupation: waiter/cook
- Occupation: watchman
- Occupation: other
- Number of working days (per week)
- Poverty scorecard PCA, based on principal components analysis of: female head/spouse education level score, number of irons owned score, number of towels owned score, number of mosquito nets score, number of household member score, number of habitable rooms score, amount of lighting fuel score, male head of house occupation score, and floor material score[8]
- Poverty scorecard PCA above 25th percentile

[8]Specifically, we developed a score based on the following point assignments: female head/spouse education: 11 points for secondary form 4, some college, completed college/graduate, or completed undergraduate/graduate degrees; 6 points for primary standard 8 or secondary forms 1–3; 4 points for primary standard 7; 1 point for primary standards 1–6; and 0 points for none/pre-school. Irons owned: 4 points if any iron is owned; 0 otherwise. Frying pans: 7 points if ≥ 2; 3 points if exactly 1; 0 points if none. Towels owned: 10 points if ≥ 2; 4 points if exactly 1; 0 points if none. Mosquito nets: 4 points if ≥ 2; 2 points if exactly 1; 0 points if none. Household size: 32 points for 1–2 people; 22 for 3; 18 for 4; 12 for 5; 8 for 6; 5 for 7–8; and 0 for ≥ 9. Habitable rooms: 8 points for ≥ 4; 5 for 3; 2 for 2; 0 for 1. Lighting fuel: 12 points for electricity/solar/gas; 6 for paraffin/candles/biogas/other; 0 for firewood/grass/dry cell. Male head/spouse occupation: 12 points for agriculture, hunting, forestry, fishing, mining, or quarrying; 9 for any other occupation; 3 if no male head/spouse; and 0 if not working. Floor material: 3 points for cement/tiles; 0 for wood/earth/other.

  - Poverty scorecard PCA above 50th percentile
  - Poverty scorecard PCA above 75th percentile
  - Literacy: can read easily
  - Literacy: can write easily
  - Self-reports having a regular salary
  - When confronting technical problems, survey participant resolves such problems by going to a repair shop
  - When confronting technical problems, survey participant resolves such problems themselves
  - Self-reported technology skills: advanced or above
  - Self-reported technology skills: expert

- Behavior-based outcomes, derived from observed behavior in participants' first (control) week
  - Number of calls made during the first week
  - Number of texts sent in first week
  - Number of persons called in the first week
  - Number of persons texted in the first week
  - Usage of Facebook app (binary indicator for any use)
  - Usage of Whatsapp app (binary indicator for any use)
  - Activity PCA, based on principal components analysis of: number of calls made during the first week and number of texts sent during the first week
  - Number of friends PCA, based on principal components analysis of: number of persons called in the first week and number of persons texted in the first week
  - Activity and Facebook/Whatsapp Usage PCA, based on principal components analysis of: number of calls made during the first week, number of texts sent during the first week, Facebook usage (binary indicator for any usage), and Whatsapp usage (binary indicator for any usage)
  - Number of friends + Facebook/Whatsapp Usage PCA, based on principal components analysis of number of persons called in the first week, number of persons texted in the first week, Facebook usage (binary indicator for any usage), and Whatsapp usage (binary indicator for any usage)
  - First-week phone usage at night (if screen was on at any point between 12am and 5am)
  - Number of distinct locations visited in the first week (distinct locations defined as number of longitude/latitude clusters in optimal K-means clustering, with a maximum of 10; see below for further details)
  - Number of private wifi networks connected to in the first week (for which the user is the only one in our sample that connected to a network of that name)
  - Number of semiprivate wifi networks connected to in the first week (for which there is no more than one other person in our sample that connected to a network of that name)

S2.6.2 Candidate features

The Sensing app captured a total of $\bar{K}$ features representing user behaviors, including:

- **App Usage:** includes binary variables for whether an app is installed (newly or at all), total number of new installs over the preceding week, the number of app session observations, the number of non-system app session observations, the number of unique apps used, usage of each app (binary for any use during the week, weekly minutes, number of days, mean daily minutes, and standard deviation of daily minutes). These statistics are calculated per app and per category of app. We also compute how many apps in a given category are used.[9]

- **Mobility:** the app attempts to capture the physical location of the phone every 5 minutes, but may be less frequent if a phone is off, or out of service range. These data are then processed into participant-week level statistics of mean, maximum, and minimum latitude and longitude; the distance from the mode coordinates to the Busara offices; and the fraction of time recorded within a 2/5/10 kilometer distance of the Busara offices. Additionally, an algorithm is used to deduce the number of 'important locations' visited in each participant-week (up to a maximum of 10) by evaluating the optimal number of k-means clusters based on silhouette score for the user's weekly longitude and latitude data, where silhouette score is computed using the conventional formula of $\frac{a-b}{max(a,b)}$, where a is the mean intra-cluster distance and b is the mean nearest-cluster distance.

- **Battery:** every 15 minutes, the app captures battery levels, so long as the phone is on and the app is active. These data are then processed into participant-week level statistics of the number of unique power sources used; the fraction of charge time on AC or USB (respectively) out of total charge time (defined as AC+USB charge time); the mean, max, min, and standard deviation of battery level observations; the most common battery health status; total minutes of charge time, total minutes of charge time at night, and fraction of time charging on AC at night, computed as time charging on AC at night divided by time charging on AC or USB at night.

- **Calls:** for each call, the app captures duration, time of day, and an anonymous identifier for the other party. These are summarized into participant-week level statistics of the

[9]We use the category definitions from the Google Play app store, plus a category for gambling apps that is constructed based on an analysis of app characteristics, and a composite category of all apps.

number of calls; the mean, min, max, and standard deviation of daily calls (overall, restricted to contacts, or restricted to noncontacts); the mean, min, max, and standard deviation of call duration; the number of unique phone numbers interacted with; the mean, min, max and standard deviation of the daily number of unique phone numbers interacted with daily; and the fraction of calls that are between participants in the study. These statistics are also calculated for various subsets of calls, defined by calls of particular durations (missed calls, calls that last at least 0, 30, 60, or 300 minutes), times (during the workday, a weekday, weekend, evening, early morning), and direction (incoming or outgoing).

- **Text messages:** the app captures the time of each incoming and outgoing text message. These raw data are then processed into participant-week level statistics: the number of texts; the mean, min, max and standard deviation of daily text messages; the number of unique phone numbers interacted with; and the mean, min, max, and standard deviation of the daily number of unique phone numbers that the user interacts with. These statistics are also calculated for various subsets, defined by times (during the workday, a weekday, weekend, evening, early morning), and direction (incoming or outgoing).

- **Wifi networks:** the app captures usage of wifi hotspots, which are summarized into participant-week level statistics of the number of weekly Wifi event observations; the number of Wifi hotspots as well as the number of Wifi hotspots with no other users connected at the time or at most one other user connected at the time; and the maximum number of Wifi networks detected by the phone at any given moment.

- **Screen:** the app captures the number of times the screen is turned on or off, the total amount of time a screen is on and off, and the maximum single span of time that a screen is on or off (calculated over the week, and for two subsets: late-night periods, and workdays).

S2.6.3 Selecting the subset of behaviors to measure costs for ($K \subset \bar{K}$)

From these $\bar{K}$ features we focused on a subset K on which to estimate costs. We first omitted features with low variance ($\geq 95\%$ of observations equal to the median).

We then identified features that were predictive of outcomes by running LASSO regressions of each candidate outcome y_i^{unscaled} on behavioral features, using person-weeks from control observations (where $\mathbf{x}_{it}(\boldsymbol{\beta} = \mathbf{0})$). We considered features that were chosen by LASSO for at

least one outcome for inclusion in K, and also augmented this candidate set with 'alternative' features, defined according to three criteria:

- We considered all variables that were defined over a similar realm of behavior. For example, if a chosen indicator was the weekly number of outgoing texts at night, we considered other indicators that relate to outgoing texts at night: the number of unique phone numbers texted at night, the maximum daily phone numbers texted at night, etc. We also considered similar measures for opposing directions (for example, incoming texts, which may be harder to manipulate), and substitutes (such as WhatsApp usage). If the indicator had to do with a certain app, we considered behaviors associated with use of that or related apps.

- We considered other features that were highly correlated (Pearson correlation coefficient ≥ 0.75) with a chosen feature, since such features were likely to capture similar behaviors.

- We considered all features that were selected from a LASSO regression that excluded from the covariate set binary features for which there was a numeric related feature (i.e., it excludes whether a given app is installed, which has numeric counterparts such as number of days the app is used) and excluded all features we deemed hard to describe.[10]

After constructing this set of alternatives, we evaluated the performance (R^2) of 3-variable LASSO models based on these alternative features relative to 3-variable LASSO models with the chosen feature. We also evaluated the performance of a single-variable model using just the alternative feature. We did this for each alternate feature, and selected several such features that we expected would have different manipulation costs from the chosen features, but which were nonetheless predictive.

Of the features we incentivized, we omitted a small number from cost estimation. Some were assigned only in the last weeks of Phase 0 and 1, and as a result we had not received data in time to include them when estimating the costs needed for Phase 2. Additionally, for a small number of features, data was missing from many participants so that costs were not identified net of fixed effects.[11] Figure S3 shows the correlation between the K features ultimately included for cost estimation.

[10]We considered a feature hard to describe if it included measures based on standard deviations, related to missingness, or was very precise in terms of timing or type of specific behavior (such as the number of outgoing calls of at least 30 minutes duration that occur between 9am and 5pm on weekdays).

[11]This underidentification was established after preliminary costs estimation, and so a small number of features were included in preliminary costs estimation but not final costs estimation.

S2.6.4 Selection of outcomes

At the end of Phase 1, we then estimated both naïve and strategy-robust decision rules for each outcome, using the set of K behavioral features for which we had cost measurements for.

Outcome labels were created by transforming the unscaled variables. For binary variables, we simply multiplied by 1000, so that each $y_i \in \{0, 1000\}$. For continuous variables, we transformed the unscaled variables with the formula:

$$y_i = \max\left\{0,\ \min\left\{1000,\ \frac{1000}{Q_{90}(y_i^{\text{unscaled}})}\, y_i^{\text{unscaled}}\right\}\right\}.$$

where $Q_{90}(y_i^{unscaled})$ represents the 90th percentile of that variable across individuals.[12]

We then selected focal outcomes for Phase 2 based on a set of heuristic criteria:

1. outcomes that would be intuitive to describe to participants
2. outcomes that could be explained by decision rules with relatively high out-of-sample predictive power
3. outcomes that we believe would not lead to potentially unhealthy behavior out of ethical concerns (such as gambling), or that might be deemed culturally inappropriate (such as gender)
4. outcomes for which naïve and strategy-robust approaches resulted in sufficiently different decision rules (in terms of coefficient selection, shrinkage, or predicted performance under manipulation)

As Phase 2 went on, we restricted consideration further to outcomes that had not yet been exhaustively assigned in preceding Phase 2 weeks. Note that the last heuristic criterion will tend to select outcomes described by decision rules that are more sensitive to manipulation and may not necessarily be representative of a random draw of outcomes.

S2.7 Deviations from pre-analysis plan

Our design was pre-specified in a pre-analysis plan (PAP) registered in the AEA RCT registry under AEARCTR-0004649. Our experiment differed from the PAP in a few ways:

[12]Note that this scaling formula inadvertently truncated outcomes with some negative values at 0, which affected two implemented outcomes (activity PCA and number of friends PCA). Because this formula was consistently applied to define our outcomes, our results remain internally consistent.

- Our budget increased, which allowed us to increase the size of the top-up sample in October 2019. The onboarding period for these top-up participants took concomitantly longer. Since we onboarded more top-up participants than anticipated, not all top-up participants were pre-recruited as part of a 'master list' as described in the PAP.

- To accommodate this longer onboarding period, we combined our planned 'Phase 0' and 'Phase 1' into a single 'Phase 0/1'. That is, instead of having all individuals simultaneously assigned 'control' challenges before assigning anyone 'incentivized' challenges, we instead assigned 1/3 of our sample every week into 'control' challenges and 2/3 every week into active 'incentive' challenges. Participants were always assigned a 'control' challenge in their first week in the sample.

- We increased the level of incentives to top out at 1000 Ksh. instead of 250 Ksh. in late October 2019, also because of our increased budget.

- We extended Phase 2 for four additional weeks, and correspondingly assigned additional outcomes, because of technical implementation issues in the last weeks of 2019 (described in detail below), and because our increased budget made this possible.

- We added a small set of potential Phase 2 outcomes to our consideration set that were based on observed activity in participants' first week in the study to capture ground truth behavioral data without any bias arising from self-reporting, in addition to the survey-based outcomes that we outline in the PAP. When selecting outcomes for Phase 2, the PAP listed the first two heuristic criteria (1-2: sufficiently high predictive performance and that the outcome could be explained intuitively). In the experiment, we also considered the two additional criteria (3-4: avoiding outcomes that could have raised ethics concerns, and whether the naïve and strategy-robust decision rules were sufficiently different).

- We used a correlation threshold of 0.75, instead of 0.9, to evaluate potential alternative features for cost measurement in Phase 1.

- We planned to evaluate decision rules using mean squared error (MSE). We instead report a monotonic transformation of that number, the square root of MSE (RMSE), because it is denominated in more natural units (currency).

- The hypothetical cost surveys were delayed until after the app data was collected because our team was focused on logistics. Because of this, we did not have the

opportunity to administer the survey to Busara participants as the study was going on, and were only able to administer the study to Mechanical Turk respondents and topical experts. We also refined the survey questions after piloting; see Section S3 for details.

S2.7.1 Technical implementation issues

In addition to the experimental deviations described above, there were a few technical issues with the Sensing app, which required us to drop certain weeks of data from our main analysis.

- Issues with Busara's data server made it difficult for some of the participants to upload data during the first two weeks of Phase 2 of the experiment.[13] For participants who were able to upload, we still have complete data, so we include these weeks in our 'all weeks' sample. In addition, we extended our planned Phase 2 by 4 weeks, to cover the first 4 weeks of January. Participants were informed of the extension over the holiday break and invited to continue. The size of the sample remained roughly constant over the break, with 970 participants remaining 'active' in the study for the first week of January, compared to 984 in the last week of December.

- In Week 7 of Phase 2, there was a multi-day data blackout for the Busara server, and behavior during the middle of the week was not recorded. Since we cannot recover complete behavior information for any individual during that week, we exclude that week from all analysis.

- One of the outcomes assigned in the fourth week of January (below-median age) was miscoded, due to a change in how year-of-birth was formatted in the survey. Roughly 90% of participants had an erroneously missing year-of-birth, which prevented us from using that outcome.

Table S3 shows that our method continues to outperform standard machine learning methods when including these weeks. More detailed results, broken down by each individual outcome, are available at `https://dan.bjorkegren.com/manipulation-appendix-extra.pdf`

[13] A power outage prevented many participants from uploading their data to the server. Later, we learned that the app had a previously undiscovered flaw by which data were not uploading to the server. Fixing this issue required addressing the problem phone by phone.

S2.7.2 Changes to method

During Phase 2 and after the completion of the experiment, we made a number of updates to our structural model. Two fixed errors; some updated parameters; and others made the method more elegant. As a result, the decision rule $\boldsymbol{\beta}^{SR}$ that was implemented in some weeks slightly differed from the one that would be estimated with the final structural model we present in this paper ($\boldsymbol{\beta}^{SR_{final}}$). We first show that these differences are minor: implemented decision rules $\boldsymbol{\beta}^{SR}$ are very close to the final $\boldsymbol{\beta}^{SR_{final}}$ in both the final likelihoods they attain, and in coefficient space. We then discuss each difference in detail.

Implemented decision rules were similar to those derived from our final model. We first measure fidelity of implemented decision rules to the final loss function, as an omnibus test. In the paper (Section 4.3.1), we define the loss improvement of decision rule $\boldsymbol{\beta}$ as $\Delta L(\boldsymbol{\beta}) = \frac{L_{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^{\text{naïve}}) - L_{\boldsymbol{C}}(\boldsymbol{\beta})}{L_{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^{\text{naïve}}) - L_{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^{SR_{final}})}$. This reveals the improvement in loss that $\boldsymbol{\beta}$ is expected to attain relative to the naïve decision rule, relative to the gain attained by the final SR decision rule $\hat{\boldsymbol{\beta}}^{SR_{final}}$, all measured according to the final loss function (the objective used in equation (6)). Figure S5 shows this value by outcome and week.

Even though implemented SR decision rules differ in some ways from the decision rules resulting from our final estimates, nearly all implemented SR decision rules are expected to improve on naïve performance and achieve close to the performance of the optimal decision rule per the final loss function. The median decision rule achieves 90% of the improvement of the optimal SR decision rule. We also observe that fidelity does not change substantially over the course of the experiment. Our main results thus include all implemented weeks for which we have data. In one outlier in week 2020-1, the optimal decision rule would be expected to perform worse than the naïve decision rule, but in this case the optimal and naïve rules are expected to perform nearly identically: the extreme relative difference is driven by a small absolute difference. The slightly higher loss of the optimal decision rule could be driven by sampling or noise.[14] A more detailed breakdown that illuminates this case is shown in Figure S6.

Second, we compare the decision rules themselves. The implemented SR decision rules $\boldsymbol{\beta}^{SR}$ themselves are also similar to the optimal under the final model, $\boldsymbol{\beta}^{SR_{final}}$. Table S5 compares the decision rules implemented in the experiment to those that would be suggested by the final cost estimates for our main outcome. Corresponding tables for all outcomes are

[14] For the outcome Advanced Technology Skills, the optimal decision rule is expected to achieve RMSE with penalization 4.98, versus 4.99 for the naïve decision rule.

available at `https://dan.bjorkegren.com/manipulation-appendix-extra.pdf`. In most cases, the final procedure selects similar values for similar coefficients.

Third, we compare the underlying cost estimates. Table S4 shows the 'preliminary cost estimates', which were calculated in week 2020-1, for all estimated behaviors. These preliminary cost estimates are similar to the final cost estimates presented in Table A1 of the paper.

Because the behavior we observe was influenced by the decision rules that were actually implemented, the decision rules we report in the main paper are those that were implemented (derived from the cost estimates at the time). However, the main paper presents the final procedure, final cost estimates, and predicted error computed using the final cost estimates. As noted above, the differences between the implemented method and the final one are minor.

In Online Appendix Section S4.5 we compare the performance of implemented decision rules $\boldsymbol{\beta}$ to what would be predicted by various models of behavior. That analysis considers how well the behavioral response to a given decision rule $\boldsymbol{\beta}$ is described by different models of manipulation. Because that analysis does not rely on whether an implemented decision rule is optimal, it provides an independent check of the model.

Differences in implemented decision rules. We now detail differences in the decision rules implemented week to week, and the final decision rule we estimated for the paper. Differences are summarized in Table S6.

One difference arose from data:

- There was no pause between Phases 1 and 2, and it took time to process the incoming data from Phase 1. Decision rules for the first four weeks of Phase 2 were estimated on a sample that omitted the final two weeks of Phase 1. During the holiday break before 2020, we updated these raw data files and re-estimated our cost matrix, so that the remainder of decision rules were estimated on the full Phase 1 data.

Two changes corrected mistakes:

- After the experiment completed, we had time to more carefully check the optimality conditions and found that the costs estimated during the experiment were a local optimum. We also had time to improve the optimization procedure (by searching more start points and implementing our own optimization algorithm), which identified a better optimum.

- Early decision rules were estimated using slightly different distributions of gaming ability than suggested by the model. This difference occurred in how draws from the distribution of unobserved gaming ability V were used in the training of decision rules. The model in the paper suggests that the random draws of unobserved gaming ability, $\tilde{v}_i$, should enter individual gaming ability as $\tilde{\gamma}_i = \frac{1}{e^{\boldsymbol{\omega}'\mathbf{z}_i + \tilde{v}_i}}$. Until the last week of Phase 2, decision rules instead mistakenly computed these random draws as entering outside of the denominator, $\tilde{\gamma}_i = \frac{1}{e^{\boldsymbol{\omega}'\mathbf{z}_i}} + \tilde{v}_i$. This results in different idiosyncratic distributions of $\tilde{\gamma}_i$.

  We investigate this in Figure S7. We plot the two cumulative distributions of $\tilde{\gamma}_i$: $\frac{1}{\gamma} + \tilde{v}_i$ (labeled as 'incorrect') versus $\frac{1}{\gamma + \tilde{v}_i}$ (labeled as 'correct') using the preliminary distribution $\tilde{V}$ that was the basis for our implemented decision rules, for both $Z = 1$ and $Z = 0$. Overall, visible differences are small, but nonetheless a Kolmogorov–Smirnov test of the equality of the two combined distributions (combining both $Z = 1$ and $Z = 0$, and comparing the different heterogeneity terms) finds a significant difference between the full distributions with a p-value of 0.018. Separate Kolmogorov-Smirnov tests for $Z = 1$ and $Z = 0$ lead to p-values of 0.403 and 0.041, respectively, significantly different for $Z = 0$ but not for $Z = 1$.

  Overall, our method is compatible with many methods of recovering a distribution of unobserved gaming ability, and that distribution is also controlled by hyperparameter ϕ, which appears to have more influence than this difference.

Some weeks used different modeling choices:

- The analyst may choose the covariates that define observable heterogeneity ($\mathbf{z}_i$). In our final model and during all weeks but the 4th week, we allowed $\mathbf{z}_i$ to be whether a person self reports their technical skills as 'advanced'. In Week 4, we instead allowed $\mathbf{z}_i$ to vary by education (education primary school or lower).

- The level of shrinkage of unobserved gaming ability, ϕ, is also a choice. Prior to observing any Phase 2 data, we used simulations to assess the decision rules that would result under different levels of ϕ. We found that $\phi = 0.005$ tended to change the selection of covariates included in the SR decision rule relative to the naïve decision rule, and was also predicted to yield a difference in performance. After we began observing Phase 2 data, we monitored how well realized fit was matched by the behavioral model under different levels of ϕ. Based on this, we opted not to change ϕ, with one exception: in week 6 we tried setting $\phi = 0.01$.

The final costs implied a different distribution of unobserved gaming ability, so for the final decision rule in the paper, we took a range of values of ϕ and evaluated the L2 distance between the resulting coefficients and those of the implemented decision rule for our final week outcomes of income and intelligence (based on Ravens). We selected the value that minimizes this distance with the respective implemented decision rules, which yields final value $\phi = 10^{-6}$. We assess sensitivity to ϕ in Online Appendix Section S4.5.

We also made minor changes to make the method more elegant and easier to describe:

- We slightly tweaked moment conditions to make them more elegant after 2019. This induced minor changes on how errors were weighed (see Appendix A1.1 for the final moment conditions):
  - Our later models use the moment condition $\mathbb{E}[\beta_{itk}\eta_{itj}] = 0$. In the first few weeks of the experiment we used a version of this condition divided by β_{itk}^2: $\mathbb{E}[\frac{\eta_{itj}}{\beta_{itk}}] = 0$. The change made it more similar to other estimators.
  - Our later models use the heterogeneity moment condition $\mathbb{E}[z_{li} \cdot v_i] = 0$. For the first few weeks we used a variant that multiplied each element in the interior sum by $\frac{1}{c_{kk}}$. The former is simpler to describe.
  - Our later models include an empirical moment condition forcing the idiosyncratic error to be mean zero ($\mathbb{E}[v_i] = 0$); early weeks did not include this moment (they only included the orthogonality condition).
- For the final decision rules, we refined the regularization protocol to select penalization closer to the boundary of 3 coefficients.[15]
- For the final decision rules, we changed the LASSO sample to coincide with that used for the SR model (it includes only individuals with non-missing tech skills, dropping approximately 1.5 percent of the sample).

S3 Can surveys measure manipulation costs?

We additionally evaluate how well hypothetical questions can predict the costs of manipulating different behaviors, using a method similar to DellaVigna and Pope (2016).

[15] Previously we relied on a grid search across 100 points, defined using the 'glmnet' R module default. We refined this in our final model to perform binary search (tol=1e-5) to recover the boundary value between a 3 coefficient decision rule and a >3 coefficient rule and use that as our penalization parameter.

S3.1 Intake

We sent a survey to two groups of participants:

- **mTurk participants** were recruited via Amazon Mechanical Turk's platform, with the posting, '*Choice Survey (15 minutes): This is a survey about choices, for smartphone owners.*' To be eligible, participants had to have successfully completed more than 100 tasks on the platform, with an overall approval rate above 95%. Participants were paid a base rate of $2 for completing the survey plus an incentive payment of $0.20 for each prediction that was within 25% of what we found in the real-world experiment.

- **Topical experts** were recruited based on professional work in a relevant discipline (development economics, machine learning, computer science), from the authors' professional networks, and from Busara staff who had not worked on this project.[16] The expert whose predictions were closest those from the real-world received a $50 prize.

S3.2 Survey design

After obtaining consent and displaying an initial screen with information, we asked subjects to predict how Kenyan smartphone owners would use their smartphones under different conditions. For example, one question notes, 'In a typical week, Paul usually makes 3 calls to a phone number in Uganda. In one week when Paul is participating in the study, Paul is paid $0.02 for every call made to a phone number in Uganda. If you think, in the week of the experiment, Paul would make 10 calls to Ugandan phone numbers, you should write the number "10".'

We then asked respondents to predict the behavior of a typical person when incentivized, with questions of the form, 'A typical person in a typical week *sends 55 text messages*. One week during the study, people were paid *$0.02 for every text message sent*. During the week they are getting paid, how many of this action would the typical person take?' We asked this for the 13 behaviors that were selected by our empirical decision rules (those listed in Table

[16] For details on recruitment see `https://dan.bjorkegren.com/manipulation-appendix-extra.pdf`. Our respondents had a variety of expertise. 71% ranked their level of skill and familiarity with digital technology (such as computers and phones) as 4 or above, on a scale of 1-5, with 1 being a total beginner, and 5 being an expert. 71% had a 4 year college degree or above. 29% of respondents have used a prepaid mobile phone in a developing country, and 17% had spent time in Kenya. Among the 13 experts, 13 had used linear regression, 9 had used other machine learning methods (such as LASSO, Random Forest, or Neural Networks), 12 had constructed an economic model of behavior, 9 had used data science methods, and 3 had done feature engineering.

4 in the paper) plus 5 randomly selected behaviors. We also asked participants to predict whether technology experts would adjust behavior more or less than non-experts.

S3.3 Results

We use responses from 13 experts and 158 mTurkers.[17] For each behavior k we back out the implied structural costs from how much respondents predict it will respond to incentives:

$$\hat{c}_{kk} = \frac{1}{N_{survey}} \sum_r \frac{\bar{\gamma} \cdot \beta_k}{\max(0.001, \Delta_{kkr})}$$

where $\beta_k > 0$ is the incentive for behavior k, $\Delta_{kkr} = x_k^r(\beta_k) - \underline{x}_k$ represents respondent's r's prediction of the typical user's change in behavior k when incentivized by β_k, and N_{survey} is the number of respondents. We take the maximum of 0.001 and Δ_{kkr} to avoid any divide by zero errors. We set heterogeneity $\bar{\gamma} = 1$ to focus on the typical user. Results are shown in Appendix Figure A1.

When estimating decision rules with costs from these surveys, we also need estimates of manipulation noise. Although we did not ask respondents to predict this, to form some estimate we use variation in participants' responses.[18] In particular, we compute each residual $\gamma_{rk} = \frac{\Delta_{kkr} \cdot \hat{c}_{kk}}{\beta_k}$ for each respondent and behavior. We draw a random sample of these residuals of length N corresponding to the size of our main experiment sample (with replacement), fit a normal distribution with MLE to smooth it, and draw N values of γ from that distribution (replacing any negative draws with zeros to ensure that gaming ability is nonnegative).

Note that our empirical exercise includes no observed heterogeneity, so only this variance was used to calibrate random effects. Although we focus on the typical user, respondents were also asked to qualitatively predict heterogeneity in gaming ability. 58% of respondents predicted that technology experts would adjust their behavior more than non-experts; and 25% by about the same.[19]

[17] We presented our method at academic conferences starting in the fall of 2018, prior to the experiment. For experts, we asked if they had seen the project presented. 11 had seen it presented; to avoid contamination we omit their responses. We also omit one behavior that was initially misworded in the survey (number of outgoing weekday SMS). For each behavior, the survey reported a typical usage level in a typical week, and asked respondents to predict the counterfactual level. We updated these typical levels based on updated measures after the first few responses on February 17, 2020. Since we compute the change in behavior from the typical level, this should not have a major impact on results.

[18] This is likely to differ from the variation they would predict in behavior, if their responses consider the average response.

[19] The responses to this multiple choice question were ordered randomly in either increasing or decreasing order, to eliminate potential bias from order.

S3.4 Validation

The design included several checks to ensure that respondents took the survey seriously. First, after reading the instructions, participants responded to two simple questions to validate understanding of the study. To complete the study, participants had to respond correctly. Second, responses were incentivized. Third, in the final demographic survey, mTurk 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. 84% of mTurk respondents agreed with the first and last statement, and disagreed with the middle; 56% did so strongly.

Although we did not mention manipulation or gaming during the survey, several mTurk respondents used the comment box at the conclusion of the survey to indicate that they would expect such manipulation of behavior:

- 'Some of the scenarios seemed easily exploitable for hundreds of dollars e.g. if calls cost me 4 cents per minute but I'm being paid 2 cents per *second*, or $1.20 per minute, I'm just going to call a friend and we put our phones down with the line open for the entire day.'

- 'Hard to guess this, I think it would depend on how many people they know or random numbers they could manage to txt or call.'

- 'I feel like those were actually really difficult to predict. In some of the scenarios, they could actually make money by placing calls. In a third world country, I would expect them to treat it like a second job in those cases. The numbers could be anything.'

- 'I figured most people would "goose" their time on calls in order to attain more of the incentives'

S4 Supplemental results

S4.1 Testing model assumptions

S4.1.1 Functional form of costs

Under quadratic manipulation costs, the model predicts that behavior moves linearly with the scale of the incentive. We can assess this functional form assumption using the variation induced by our experiment.

In our simple challenges, the decision rule $\boldsymbol{\beta}_{it}$ randomly assigned the level of the incentive dimension k, β_{itk}, from a menu that included a variety of positive incentives (i.e., 'earn β_{itk} for each behavior k you do') and a small number of negative incentives ('lose β_{itk} for each behavior k you do'). The scale was chosen differently for each behavior k. We pool results across behaviors k, and seek to characterize the nonparametric function $h(\cdot)$ from the equation:

$$x^*_{itk}(\boldsymbol{\beta}) = \psi_k \cdot h\left(\frac{1}{\xi_k}\beta_k\right) + \theta_{ik} + \mu_{tk} + \eta_{itk}$$

where ψ_k and ξ_k are scaling factors. We set $\xi_k = \max_{it}\{|\beta_{itk}|\}$ so that $\frac{\beta_k}{\xi_k} \in [-1, 1]$ represents a standardized index of the strength of the incentive in $\boldsymbol{\beta}$. For control person-weeks where no incentives are given ($\boldsymbol{\beta}_{it} = \mathbf{0}$), we include observations from all behaviors. In person-weeks where a behavior k was incentivized ($\boldsymbol{\beta}_{it} = \beta_{itk}\boldsymbol{\delta}_k$), we only include the observation of that behavior k, omitting observations of all other behaviors. (We also omit person-weeks that incentivized more than one behavior simultaneously.) We also filter to include only behaviors k that were assigned both positive and negative incentives during the experiment, to ensure balance.

We estimate $h(\cdot)$ in four steps. First, we restrict the sample to control weeks and estimate the OLS regression

$$\mathbf{x}_{it} = \boldsymbol{\theta}_i + \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}$$

to recover $\boldsymbol{\theta}$ (natural behaviors) and $\boldsymbol{\mu}$ (behavior-time fixed effects). Second, we use these fixed effects to demean the behavior in all weeks, recovering $\Delta x_{itk} = x_{itk} - \theta_{ik} - \mu_{tk}$. Third, we determine the scale of each behavior by setting ψ_k to the standard deviation in the demeaned behaviors ($\psi_k = SD_k(\Delta x_{itk})$, where the subscript on SD indicates that this is taken over individuals and weeks in the sample of relevant weeks for behavior k). Finally, we plot $\frac{1}{\psi_k}\Delta x_{itk}$ versus $\frac{1}{\xi_k}\beta_{itk}$, as well as a loess smoother, in Figure S4. To avoid overplotting, we sort observations by $\frac{1}{\xi_k}\beta_{itk}$, divide the observations into 2,000 bins, and plot the mean of

$\frac{1}{\psi_k}\Delta x_{itk}$ versus the median of $\frac{1}{\xi_k}\beta_{itk}$ within each bin (because there are many control week observations, many bins represent incentives of zero). To a first approximation, the response to incentives is linear, but some nuances are evident: respondents appear less responsive to negative incentives, but these data are sparse since we assigned fewer respondent-weeks to negative incentives. Overall, this suggests that quadratic costs are a reasonable (if imperfect) approximation in our setting.

S4.1.2 Separability of heterogeneity by behavior and person

Our parameterization of costs allows for separable heterogeneity by behavior and by individual, with the individual-level heterogeneity including manipulation noise. Here, we assess the implication that behaviors that have lower average manipulation costs will tend to face higher manipulation noise. As in Appendix S1, we consider the case where shocks have been decomposed, $\boldsymbol{\epsilon}_{it} = \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}$, and common shocks $\boldsymbol{\mu}_t$ are absorbed with time fixed effects. The equations also apply without time fixed effects, in which case one may omit $\boldsymbol{\mu}_t$ and swap $\boldsymbol{\epsilon}_{it}$ for $\boldsymbol{\eta}_{it}$.

Section 3 lays out an empirical equation relating incentives to behavior in our experiment,

$$\mathbf{x}_{it}(\boldsymbol{\beta}_{it}) = \underline{\mathbf{x}}_i + C_i^{-1}\boldsymbol{\beta}_{it} + \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}$$

In control weeks ($\boldsymbol{\beta}_{it} \equiv \mathbf{0}$), this suggests that

$$\mathbf{x}_{it}(\mathbf{0}) - \underline{\mathbf{x}}_i - \boldsymbol{\mu}_t = \boldsymbol{\eta}_{it}$$

so that we can estimate the variance of the idiosyncratic shock to each behavior as

$$Var_k(\eta_{itk}) = Var_k(x_{itk}(\mathbf{0}) - \underline{x}_{ik} - \mu_{tk})$$

where the subscript on Var indicates that this is taken over individuals and weeks in relevant weeks (control weeks) for a given behavior k.

Incentivized weeks face these idiosyncratic shocks as well as manipulation,

$$\mathbf{x}_{it}(\boldsymbol{\beta}_{it}) - \underline{\boldsymbol{x}}_i - \boldsymbol{\mu}_t = \boldsymbol{\eta}_{it} + C_i^{-1}\boldsymbol{\beta}_{it}.$$

For simple challenges that incentivize only behavior k, we have $\boldsymbol{\beta}_{it} = \beta_{it}\boldsymbol{\delta}_k$ for i in t. If we

assume the cost matrix is diagonal ($c_{jki} \equiv 0$ for $j \neq k$), then

$$x_{itk}(\boldsymbol{\beta}_{it}) - \underline{x}_{ik} - \mu_{tk} = \frac{\beta_{itk}}{c_{kki}} + \eta_{itk}$$

or

$$\frac{1}{\beta_{itk}}(x_{itk}(\boldsymbol{\beta}_{it}) - \underline{x}_{ik} - \mu_{tk}) = \frac{1}{c_{kki}} + \frac{1}{\beta_{itk}}(\eta_{itk})$$

If η_{itk} does not covary with c_{kki} across i, then taking the variance of each side yields

$$Var_k\left(\frac{1}{\beta_{itk}}(x_{itk}(\boldsymbol{\beta}_{it}) - \underline{x}_{ik} - \mu_{tk})\right) = Var_k\left(\frac{1}{c_{kki}}\right) + Var_k\left(\frac{1}{\beta_{itk}}\eta_{itk}\right)$$

which yields the variance of the inverse of manipulation costs

$$Var_k\left(\frac{1}{c_{kki}}\right) = Var_k\left(\frac{1}{\beta_{itk}}(x_{itk}(\boldsymbol{\beta}_{it}) - \underline{x}_{ik} - \mu_{tk})\right) - Var_k\left(\frac{1}{\beta_{itk}}\eta_{itk}\right).$$

Given that β_{itk} is assigned randomly, we can expand the last term,

$$Var_k\left(\frac{1}{c_{kki}}\right) = Var_k\left(\frac{1}{\beta_{itk}}(x_{itk}(\boldsymbol{\beta}_{it}) - \underline{x}_{ik} - \mu_{tk})\right) - Var_k\left(\eta_{itk}\right)\left[Var_k\left(\frac{1}{\beta_{itk}}\right) + \mathbb{E}_k\left(\frac{1}{\beta_{itk}}\right)^2\right].$$

with corresponding standard deviation $SD_k\left(\frac{1}{c_{kki}}\right) = \sqrt{Var_k\left(\frac{1}{c_{kki}}\right)}$.

Under the separability condition in Section 3.1, $c_{kki} = c_{kk}/\gamma_i$, so that

$$SD_k\left(\frac{1}{c_{kki}}\right) = \frac{1}{c_{kk}} \cdot SD_k(\gamma_i).$$

That is, across behaviors k, the standard deviation of the inverse of the manipulation cost is proportional to the inverse of the average cost of manipulation c_{kk}.

We assess this by plotting our GMM estimate of $\frac{1}{c_{kk}}$ versus $SD_k\left(\frac{1}{c_{kki}}\right)$, as derived above, in Figure S8. We see that behaviors that are easier to manipulate on average have higher heterogeneity in manipulability, and nearly follow a proportional line as suggested by separability in our model.

S4.2 Additional comparative statics

Figures S9 and S10 present additional comparative statics.

S4.3 Additional monte carlo simulations

S4.3.1 Theoretical iterated approach, cumulative

Table S7 shows a version of an iterated approach that trains on the cumulative data collected so far, rather than the previous period. This dampens oscillations but still takes several iterations to stabilize, and even then performs worse than our method.

S4.3.2 Manipulation can improve performance

Manipulation can *improve* performance, if ease of manipulation (γ_i) is correlated with the outcome (y_i). This example can be seen in the first two rows of Table S8, where manipulation improves the performance even of naïve estimators. Our method *increases* the coefficient on the manipulated behaviors to better exploit the information contained in manipulation, and thus further improves performance, as shown in the third row.

S4.4 Additional performance comparisons

As discussed in Section 5.1 of the paper, standard machine learning estimators evaluate features based on their correlation with the outcome, without accounting for the differential costs of manipulation of each feature. Here, we evaluate two common approaches.

S4.4.1 Comparison to intuitive approach (restricted LASSO)

In Table S9, we show how LASSO performs when restricted to select from the least manipulable behaviors, as measured by our experiment. This is akin to an intuition-based approach to variable selection. We allow LASSO to select only on a subset of variables that are above the 50th or 75th percentile in cost-of-manipulation, respectively.[20]

S4.4.2 Comparison to iterated LASSO

We also use our experimentally estimated model to simulate the effects of applying the iterated approach in our setting, in Figures S11 and S12.

[20]Note that these report $LASSO_{final}$ rather than $LASSO$, which include two changes: the regularization protocol was refined to select penalization closer to the boundary of 3 coefficients and the sample was changed to coincide with that used for the SR model (it includes only individuals with non-missing tech skills, dropping approximately 1.5 percent of the sample).

S4.5 Anticipating fit

Here we treat the decision rules we implemented as arbitrary, and compare the behavior they induce to what would be predicted by different models. Because this fit can be assessed for any combination of implemented decision rule $\boldsymbol{\beta}$ and model of behavior m, this exercise does not depend on whether $\boldsymbol{\beta}$ represents an optimal decision rule under m.

For each decision rule $\boldsymbol{\beta}$ we compute the actual RMSE we obtained when it was implemented, as well as the expected RMSE that would result if behavior followed $\mathbf{x}_m(\boldsymbol{\beta})$, using model m to predict the behavioral response to $\boldsymbol{\beta}$. In Figure S13, we assess variants m of models and compute the rank correlation between predicted and actual RMSE to assess how well the model captures behavior.

Standard estimators assume that behavior will remain the same as in the control weeks, which corresponds to a model of behavior $\mathbf{x}_m(\boldsymbol{\beta}) = \underline{\mathbf{x}}$. The left panel of Figure S13 shows results when performance is predicted by the standard expectation. We see that for the opaque treatment, the standard expectation does well at ranking relative performance. However, that performance drops below 50% when decision rules are transparent.

The expectations with manipulation responses use the final primitives as estimated in Section 4 to model $\mathbf{x}_m(\boldsymbol{\beta})$. We consider variants of this model that use different values of ϕ in the right panel of Figure S13.

When ϕ is very low (near zero), at the left of the right panel, manipulation noise is turned off. The model thus anticipates only deterministic shifts, and does not expect information to be muddled (Frankel and Kartik, 2019) (as described in Section 2.3). Here, results are similar to the standard model: there is little manipulation adjustment, and its ability to predict behavior in the transparent treatment is about as poor as the standard model. As ϕ increases, the model anticipates more noise in the manipulation response. When ϕ is sufficiently high, we see a reversal: the model begins to better predict behavior in the transparent treatment than the opaque treatment. That is, the model better captures the manipulation that transparency induces. At the final model setting of $\phi = 10^{-6}$ (see Online Appendix S2.7.2), we see that our model is able to anticipate the performance of an implemented decision rule when transparent almost as well as the standard expectation is, for an opaque decision rule.[21] In hindsight, it appears that a slightly smaller choice of ϕ would have achieved an even higher rank correlation

[21] A potential concern with focusing only on this ϕ that generates noise corresponding to the implemented level is that we monitored the performance of the strategy-robust adjustment as the experiment progressed. Although we did not change ϕ as the experiment progressed (with the exception of one week), one might be concerned that the selection of ϕ in later weeks was set based on the model's predictive performance in earlier weeks. This would reduce the validity of this exercise for earlier weeks, since our final ϕ would then be set to improve the fit in earlier weeks.

than the one we implemented. As ϕ increases beyond that level, the model anticipates too much manipulation noise: it still better predicts performance in the transparent treatment than the opaque treatment, but does so worse than at the optimal ϕ.

These results suggest that manipulation noise (unobserved heterogeneity) is an important component of manipulation, and that properly accounting for it can help anticipate model performance. This also demonstrates an empirical approach to calibrating ϕ: after observing behavior under any implemented decision rule $\boldsymbol{\beta}$, one can select the behavioral model with ϕ that best explains actual performance.

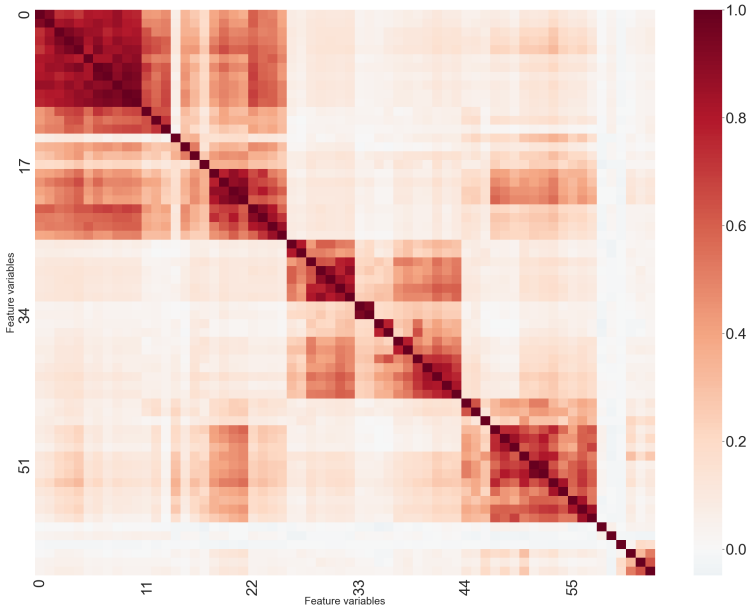
One caveat to the above results is that while we do quite well at ranking model performance, the manipulation correction is sometimes overly pessimistic about the *levels* of manipulation that naïve decision rules would encounter (in terms of absolute RMSE). That pessimism leads it to avoid certain regions of the parameter space. In contrast, we find that we can better predict the levels of manipulation of SR decision rules. This suggests that the method works not by predicting levels of manipulation exactly, but rather by avoiding areas of parameter space that are prone to manipulation. It also suggests that it may not be essential to measure manipulation costs precisely; rules may perform well as long as they can be trained to avoid problematic regions of the parameter space.

References

DellaVigna, Stefano and Devin Pope, "Predicting Experimental Results: Who Knows What?," Working Paper 22566, National Bureau of Economic Research August 2016.

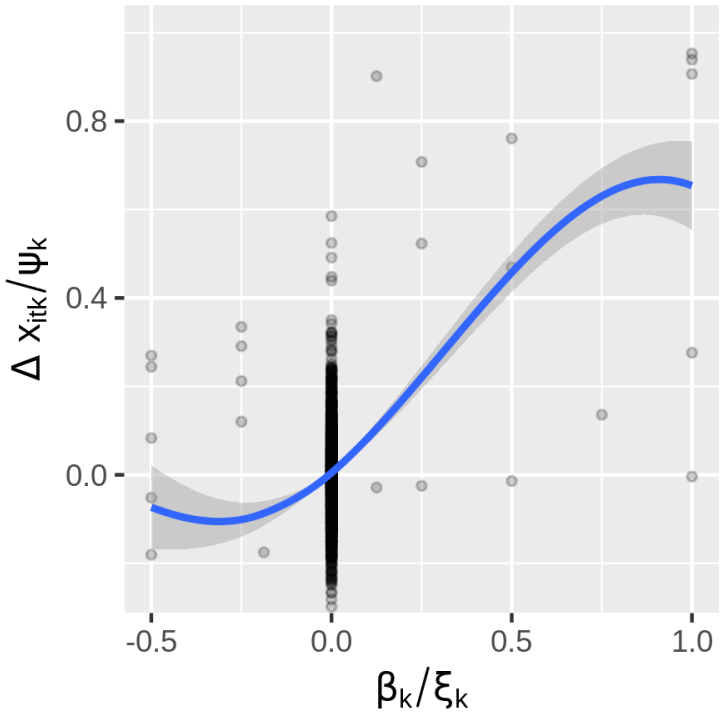
Frankel, Alex and Navin Kartik, "Muddled Information," *Journal of Political Economy*, August 2019, *127* (4), 1739–1776.

Figure S3: Correlations between Behaviors



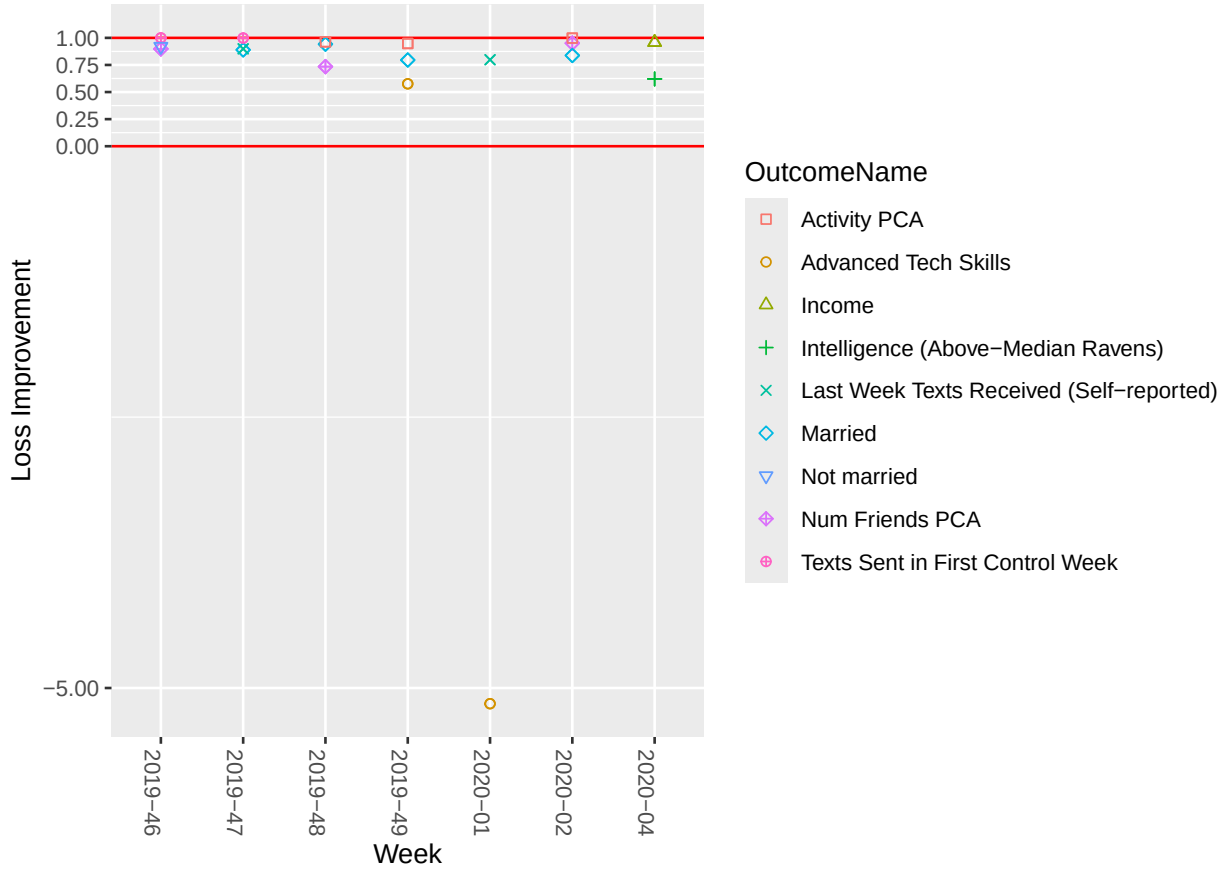
Notes: Each row and column represent a feature of behavior. Each cell indicates the correlation in activity between those two features across individuals in our study (estimated on person-level averages from control weeks). For display purposes, we use hierarchical clustering to group similar features.

Figure S4: Functional Form of Response to Costs



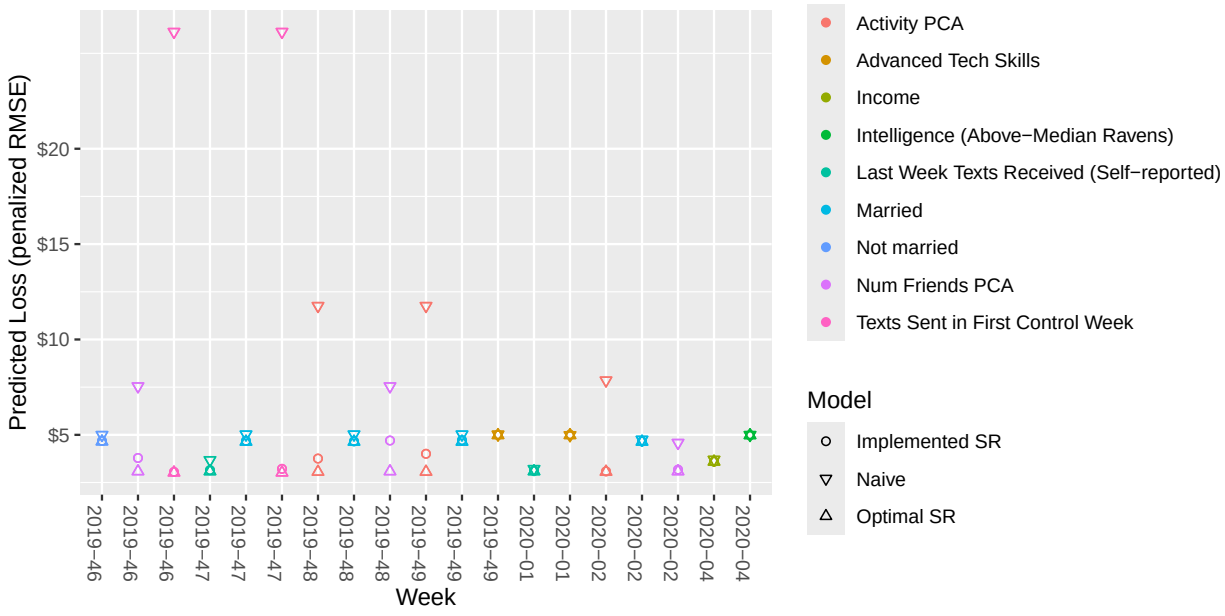
Notes: Figures show how much behaviors shift ($\frac{\Delta x_{itk}}{\psi_k}$) in response to the (randomized) strength of the financial incentives ($\frac{\beta_{itk}}{\xi_k}$). Normalization factors are set so $\psi_k = SD_k\left(\Delta x_{itk}\right)$ and $\xi_k = \max_{it}\{|\beta_{itk}|\}$. 179,475 person-week-behavior observations are binned into 2,000 points by $\frac{\beta_{itk}}{\xi_k}$, then plotted at the median $\frac{\beta_{itk}}{\xi_k}$ and mean $\frac{\Delta x_{itk}}{\psi_k}$ for that group. The large mass of points in the center are the control weeks where no incentive was given. Figure includes only behaviors that were assigned both positive and negative incentives during the experiment. We omit 88 observations with very low incentives ($\tilde{\beta} < -0.5$), as few such incentives were given in the full experiment. Blue line shows a loess fit with span parameter 1.5.

Figure S5: Loss Improvement of Implemented Relative to Final Decision Rules



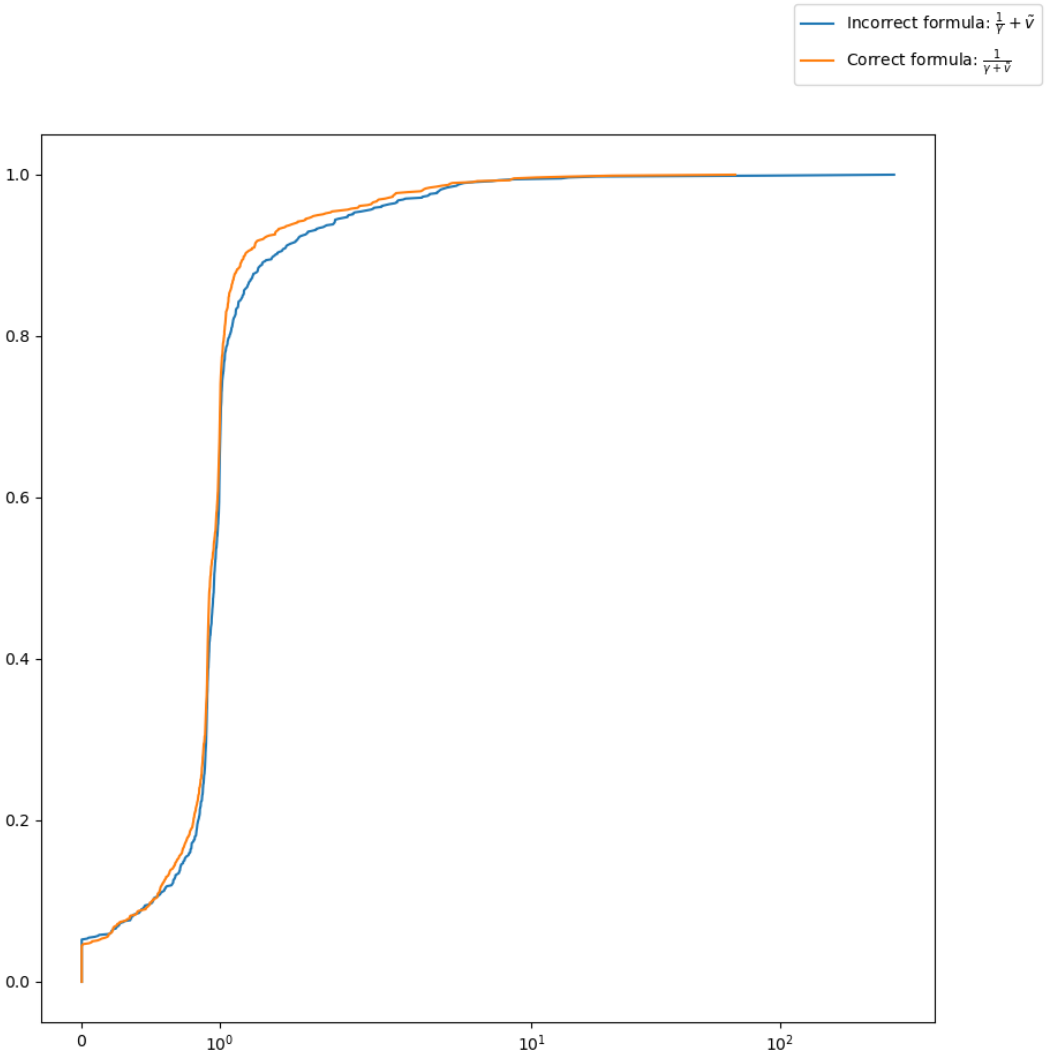
Notes: Figure shows the loss improvement $\Delta L(\boldsymbol{\beta}) = \frac{L_{\boldsymbol{C}}(\boldsymbol{\beta}^{\text{naïve}}) - L_{\boldsymbol{C}}(\boldsymbol{\beta})}{L_{\boldsymbol{C}}(\boldsymbol{\beta}^{\text{naïve}}) - L_{\boldsymbol{C}}(\boldsymbol{\beta}^{SR_{final}})}$ for implemented decision rule $\boldsymbol{\beta}$ as a proportion of that achieved by the final SR rule $\boldsymbol{\beta}^{SR_{final}}$ relative to the naïve rule $\boldsymbol{\beta}^{\text{naïve}}$. Among implemented decision rules $\boldsymbol{\beta}$, only SR decision rules are shown. Loss is measured in mean squared error, including the penalization term with the assigned hyperparameter $\lambda = \underline{\lambda}^{3var_{assigned}}$ for assigned decision rules and the final hyperparameter $\lambda = \underline{\lambda}^{3var_{final}}$ for final decision rules. Horizontal lines denote where the implemented rule is expected to achieve the same performance as the final rule ($\Delta L(\boldsymbol{\beta}) = 1$) and the naïve rule ($\Delta L(\boldsymbol{\beta}) = 0$).

Figure S6: Predicted Loss by Decision Rule



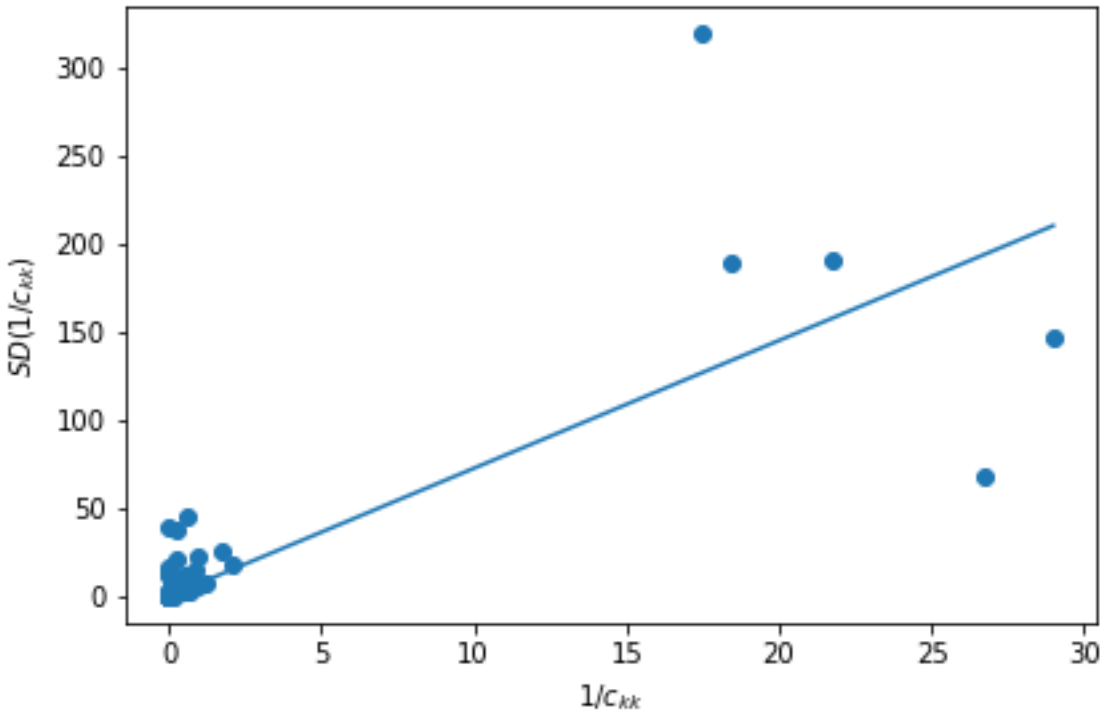
Notes: Figure shows the loss that would be expected for the implemented naïve decision rule, and both implemented and optimal SR decision rules ($\boldsymbol{\beta}^{\text{naïve}}$ and $\boldsymbol{\beta}$, respectively), as well as the corresponding final SR decision rule ($\boldsymbol{\beta}^{SR_{final}}$), for the outcomes assigned on each week. Loss is measured as the square root of expected mean squared error under the final costs plus the penalization term with the final hyperparameter $\lambda = \underline{\lambda}^{3var_{final}}$ for final decision rules and the assigned hyperparameter $\lambda = \underline{\lambda}^{3var_{assigned}}$ for assigned rules.

Figure S7: Comparison of CDFs of $\frac{1}{\gamma_i} + \tilde{v}_i$ and $\frac{1}{\gamma_i + \tilde{v}_i}$



Notes: Figure shows the cumulative distribution function of cost heterogeneity terms based on the incorrect formula ($\frac{1}{\gamma_i} + \tilde{v}_i$) that was implemented prior to January 21, 2020, alongside the cumulative distribution function of cost heterogeneity terms based on the correct formula ($\frac{1}{\gamma_i + \tilde{v}_i}$). Both CDFs are presented using the distribution $\tilde{V}_i$ based on preliminary cost estimates from 2020 week 1. Kolmogorov–Smirnov test of the equality of the two combined distributions (combining both $Z = 1$ and $Z = 0$) finds a significant difference between them at a p-value of 0.018. Separate Kolmogorov-Smirnov tests for $Z = 1$ and $Z = 0$ find p-values of 0.403 and 0.041, respectively.

Figure S8: The manipulation noise in behavior k correlates with its average manipulation level



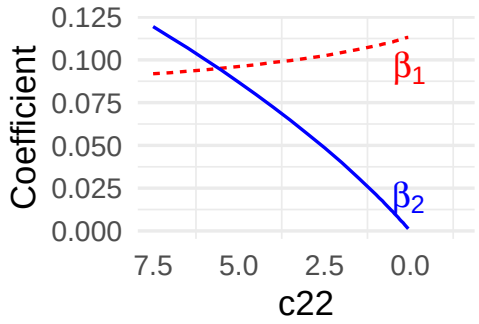
Notes: Figure shows that across behaviors, the standard deviation of manipulation noise $SD(1/c_{kki})$ is correlated with the inverse of the average manipulation cost as estimated by our model, $1/c_{kk}$. The straight line is the line of best fit through the origin, which represents the separability assumption that our model imposes. Variability in manipulation noise is computed as described in Section S4.1.2; it could not be computed for two behaviors that had more noise in the control weeks than in incentivized weeks; these are omitted from the plot.

Figure S9: Additional Comparative Statics: Cost Matrix

The first behavior is more predictive in the baseline behavior ($b_1 > b_2$), but is easily manipulable ($c_{11} \ll c_{22}$).

(a) $\boldsymbol{\beta}^{SR}(c_{22})$

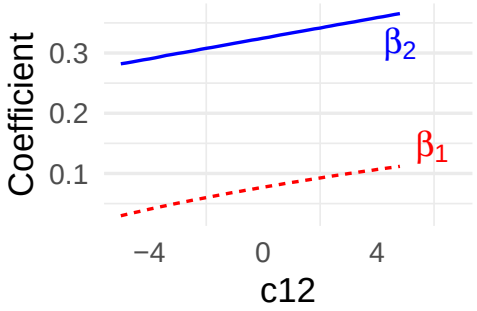
$c_{12} = 0$



As x_2 becomes cheaper to manipulate (c_{22} decreases), $\boldsymbol{\beta}^{SR}$ places less weight on it, and adjusts the weight placed on x_1.

(b) $\boldsymbol{\beta}^{SR}(c_{12})$

$c_{22} = 32$



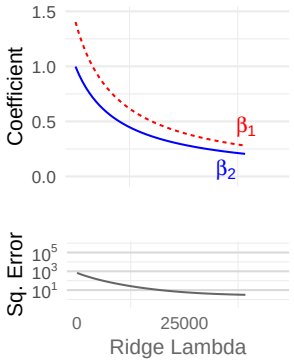
If manipulating one variable makes it easier to manipulate the other (c_{12} sufficiently negative), $\boldsymbol{\beta}^{SR}$ reduces weight on both.

$\underline{\boldsymbol{x}}_i \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, $\mathbf{b} = \begin{bmatrix} 1.4 \\ 1 \end{bmatrix}$, $C_i = \frac{1}{\gamma_i} \begin{bmatrix} 4 & c_{12} \\ c_{12} & c_{22} \end{bmatrix}$, $\gamma_i \overset{iid}{\sim} Par$ $e_i \overset{iid}{\sim} N(0, 0.25)$. Policymaker knows C but only the distribution of γ_i, when estimating the decision rule. N=10,000. Squared error measured from the same population, incentivized to that decision rule.

Figure S10: Additional Comparative Statics: Ridge and Homogeneity

C and γ_i heterogeneous The first behavior is more predictive in the baseline ($b_1 > b_2$), but is easily manipulable ($c_{11} \ll c_{22}$).

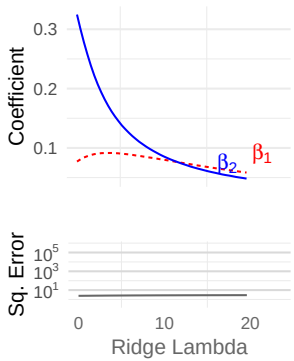
(a) $\boldsymbol{\beta}^{Ridge}$



λ ($\gamma = 1$)

Like LASSO, ridge places more weight on x_1.

(b) $\boldsymbol{\beta}^{SR}$

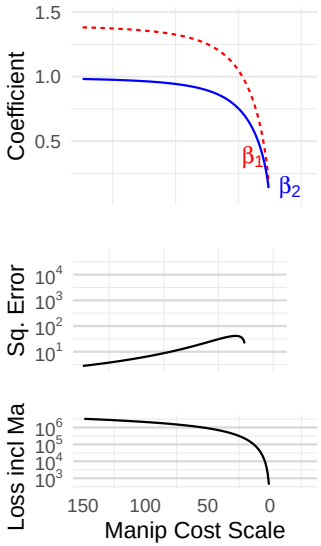


λ ($\gamma = 1$)

Our method can be combined with other forms of penalization (such as ridge shown here), to more finely manage out of sample fit.

C homogenous: Features equally costly to manipulate

(c) $\boldsymbol{\beta}^{SR}$

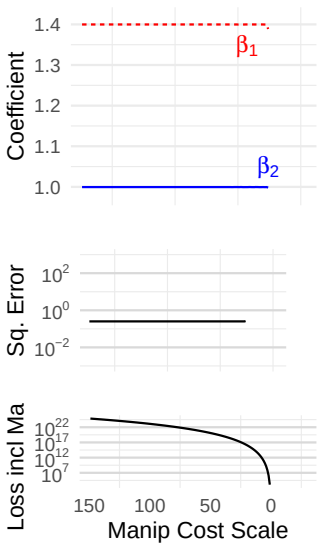


$1/\gamma$

When features are equally costly to manipulate, our method penalizes in a similar manner to ridge.

γ_i homogeneous: Same gaming ability

(d) $\boldsymbol{\beta}^{SR}$

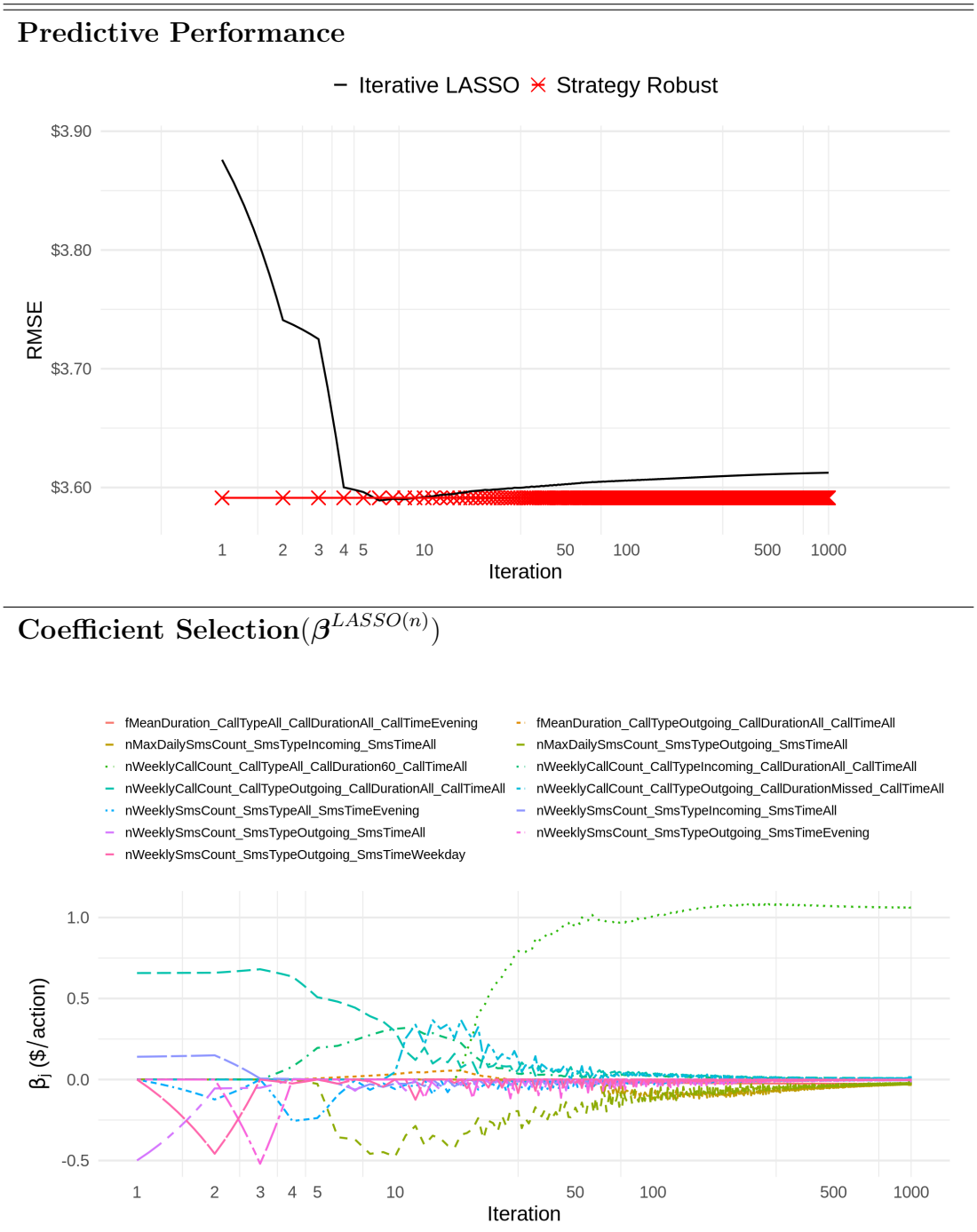


$1/\gamma$

When gaming ability is homogenous, everyone shifts behavior equally. Predictive performance does not change as manipulation costs decrease, but utility is lost on manipulation.

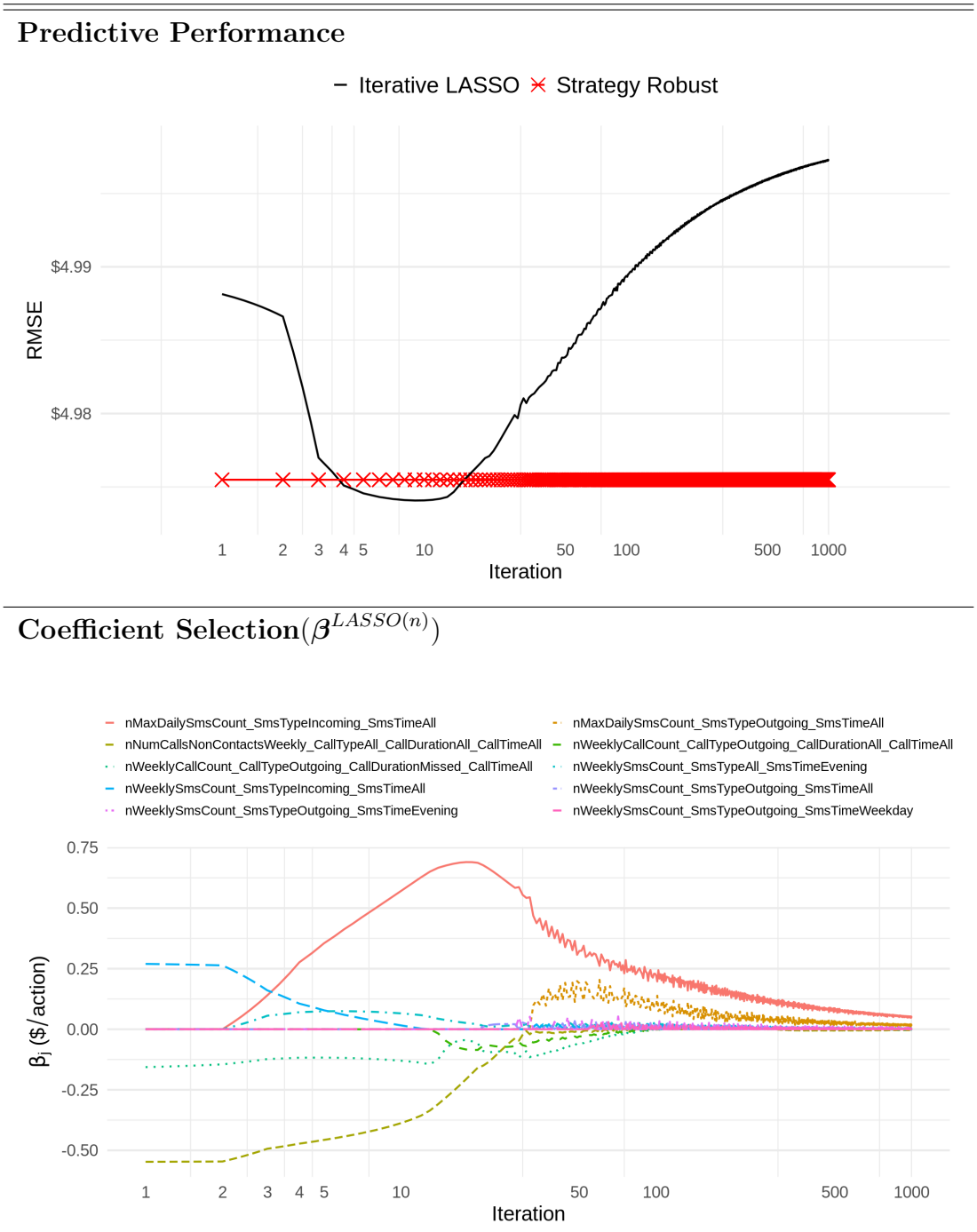
$\underline{\boldsymbol{x}}_i \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, $\mathbf{b} = \begin{bmatrix} 1.4 \\ 1 \end{bmatrix}$, $e_i \overset{iid}{\sim} N(0, 0.25)$. For (a,b,d): $C_i = \frac{1}{\gamma\gamma_i^{het}} \begin{bmatrix} 4 & 0 \\ 0 & 32 \end{bmatrix}$; for (c): $C_i = \frac{1}{\gamma\gamma_i^{het}} \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$. For (a,b,c), $\gamma_i^{het} \overset{iid}{\sim} Pareto(1, 0.1)$; for (d): $\gamma_i^{het} \equiv 0.2$. Policymaker knows C and γ but only the distribution of γ_i^{het} (which in (d) is identical for all). N=10,000, and B=10 draws of γ_i^{het}. Squared error measured on an out of sample draw from the same population, incentivized to that decision rule. 'Loss incl Ma' includes squared error plus manipulation costs: $M(\cdot) = 0.1 \cdot \sum_i \boldsymbol{\beta}' C_i^{-1} \boldsymbol{\beta}$.

Figure S11: Simulated Performance of Iterated Approach: Income



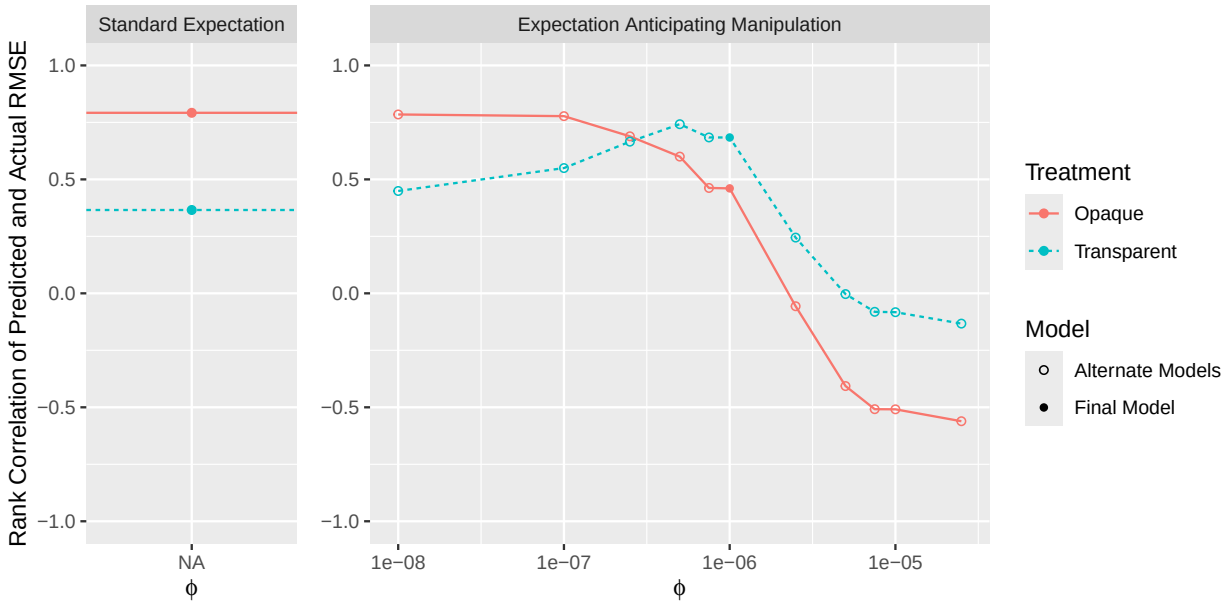
These figures show the performance of iteratively estimated LASSO for the income outcome. $\boldsymbol{\beta}^{LASSO(1)}$ is estimated using data $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta} = \mathbf{0}))$, and subsequently $\boldsymbol{\beta}^{LASSO(n)}$ is estimated using data $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta}^{LASSO(n-1)}))$ simulated from the model that was estimated with experimental data. Top panel shows performance of $\boldsymbol{\beta}^{LASSO(n)}$) on the manipulated data it would induce $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta}^{LASSO(n)}))$ for each iteration n. It also shows the performance of our strategy robust method $\boldsymbol{\beta}^{SR}$ on the manipulated data it would produce $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta}^{SR}))$. Bottom panel shows the coefficients $\boldsymbol{\beta}^{LASSO(n)}$ for each iteration n. LASSO penalization λ set to ensure 3 coefficients in $n = 1$, and held fixed as n is incremented; this sometimes results in decision rules that have a different number of coefficients.

Figure S12: Simulated Performance of Iterated Approach: Ravens (intelligence) above median



These figures show the performance of iteratively estimated LASSO for the outcome Raven's (intelligence) above median. $\boldsymbol{\beta}^{LASSO(1)}$ is estimated using data $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta} = \mathbf{0}))$, and subsequently $\boldsymbol{\beta}^{LASSO(n)}$ is estimated using data $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta}^{LASSO(n-1)}))$ simulated from the model that was estimated with experimental data. Top panel shows performance of $\boldsymbol{\beta}^{LASSO(n)}$ on the manipulated data it would induce $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta}^{LASSO(n)}))$ for each iteration n. It also shows the performance of our strategy robust method $\boldsymbol{\beta}^{SR}$ on the manipulated data it would produce $(y_i, \mathbf{x}_i^*(\boldsymbol{\beta}^{SR}))$. Bottom panel shows the coefficients $\boldsymbol{\beta}^{LASSO(n)}$ for each iteration n. LASSO penalization λ set to ensure 3 coefficients in $n = 1$, and held fixed as n is incremented; this sometimes results in decision rules that have a different number of coefficients.

Figure S13: Comparing Predicted and Implemented Performance



Notes: Figure shows the rank correlation between the actual RMSE and the RMSE predicted by various behavioral models, for the set of implemented decision rules. Separate lines are shown for treatments in implemented decision rules (transparent and opaque). Predicted performance is computed using different behavioral models along the x-axis. The left panel assumes $\mathbf{x}_m(\boldsymbol{\beta}) = \underline{\mathbf{x}}$, and uses a standard expectation $L(\boldsymbol{\beta}) = \frac{1}{N}\sum_i \left[y_i - \alpha - \boldsymbol{\beta}'\underline{\mathbf{x}}\right]^2 + R_\lambda(\boldsymbol{\beta})$. The right panel models $\mathbf{x}_m(\boldsymbol{\beta})$ as in the paper, and computes loss as the minimand of equation (6) in the paper (the strategy robust expectation), for various levels of manipulation noise as controlled by ϕ. Note that implemented decision rules were estimated using the preliminary estimates, but predicted performance is computed based on the final estimates, as described in S2.7.2.

Table S3: Performance of Decision Rules Across All Outcomes and All Weeks

	All outcomes (pooled)		All outcomes (pooled; no overlap)	
	$\boldsymbol{\beta}^{LASSO}$	$\boldsymbol{\beta}^{SR}$	$\boldsymbol{\beta}^{LASSO}$	$\boldsymbol{\beta}^{SR}$
Prediction Error	RMSE ($)			
Implemented: Opaque	3.780 (0.089)	3.710 (0.061)	5.118 (0.452)	4.406 (0.266)
Implemented: Transparent	4.641 (0.155)	4.130 (0.096)	4.682 (0.16)	4.113 (0.11)
N (Treatment person-weeks, Opaque)	1344	1344	1155	1155
N (Treatment person-weeks, Transparent)	1298	1246	1234	1162

Notes: Table reports performance of decision rules by root mean squared error (RMSE) at the week-decision-rule level, pooled across all outcomes. Outcomes include: income, intelligence (Raven's above median), activity PCA, number of friends PCA, marital status, technology skills, self-reported texts received in a single week, and observed texts sent in first control week. 'Pooled; no overlap' sample restricts evaluation sample to individuals who never received a simple challenge incentivizing the behaviors in the given decision rule. Bootstrapped standard errors in parentheses, which hold fixed the decision rule and resample individuals with replacement. Results broken down by each individual outcome are available at `https://dan.bjorkegren.com/manipulation-appendix-extra.pdf`

Table S4: Estimated Manipulation Costs for All Behaviors (Preliminary)

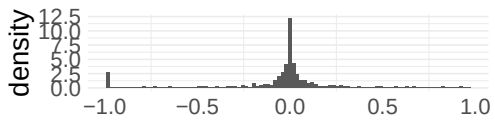
Heterogeneity by Behavior ($\mathbf{C}^{prelim}$ diagonal; all incentivized behaviors)

▲	text you send	0.03
▲ •	text you receive	0.04
	text you send on a weekday	0.05
	each text message you send in the evening hours (after 6pm)	0.05
▲ •	text you send or receive in the evening (6pm−10pm)	0.06
	call with a number not in your contacts during the workday (9am−5pm	0.58
▲ •	time you call someone	0.59
•	second of your shortest weekend call	0.64
	call with a number not in your contacts on the weekend	0.80
▲ •	each different person you text or are texted by	1.01
	person you text during the early morning hours (12am − 5am)	1.03
▲ •	time you receive a call	1.33
	person you call during the workday	1.36
	person you text on the weekday that you send the most texts	1.56
	call between 12am and 5am	1.65
	each different person you text or are texted by on the weekend	1.68
	text you send on the day that you send the most texts	1.92
▲ •	call with someone not in your contacts	2.24
	person you receive a text from during the evening (5pm − 10pm)	2.25
▲	text you receive on the day you receive the most texts	2.30
	text you send on the day you send the least texts	3.03
	if you keep it at 100%.	3.96
	each person you text in the evening (after 6pm)	4.34
•	call you make that's missed	4.64
	person you text	5.43
	person who calls you during the workday	5.96
▲	person who texts you	6.01
	person you call for at least 30 seconds during the workday (9am−5pm	6.17
	each person you call for over 30 minutes	6.61
	outgoing call on the day with the least outgoing calls	6.74
	person you text on the workday that you text the most people	7.51
▲ •	second of your average evening (6pm−10pm) call	19.01
	if the number of texts you send varies a lot weekday to weekday	26.78
	time you plug in your phone and it charges	26.98
	day you use a whatsapp−related app	31.26
	if the number of texts you send and receive varies a lot day to day	51.07
	person who calls you for at least 30 seconds during the workday (9am−5pm	89.13
	missed call	111.91
	if the number of texts you receive varies a lot day to day	130.55
	person you receive a text from during the early morning hours (12am − 5am)	149.56
	tools−type app you use	161.05
	if the number of texts you receive varies a lot weekday to weekday	193.99
	second of your average workday call (monday−friday	224.92
	text you receive on the day you receive the least texts	275.94
	5−minute−or−more call with a non−contact	318.49
	call that lasts at least an hour	365.46
	missed call on the day you have the least missed calls	444.76
	if you use a whatsapp−related app.	546.98
	day you use the app	635.14
	text you receive on the evening (6pm−10pm) that you receive the least texts	700.98
▲ •	second of your average call duration	741.46
	if you use a tools−type app.	916.65
	if you use any document− or report−related app.	1,222.66
	day you use the gmail app	1,279.22
	productivity−type app you use	1,786.24
	day you use any instagram−brand app	10,807.27
	if the number of texts you receive in the evening varies a lot day to day	11,759.37
	if you use the gmail app.	25,725.74
	day that you use a productivity−type app	25,966.47
	times the average day−to−day variation in your texts	29,917.52
	each day you use a photography app	38,940.56
	day that you use the instagram app	49,899.29
	photography−type app you use	55,959.59
	if you use any twitter−brand app.	58,866.68
	instagram−brand app used	68,081.55
	if you use a productivity−type app.	79,180.98
	if you use the app	95,576.65
	if you use the instagram app.	116,009.88
	social−type app you use	121,029.81
	if you use the twitter app at least once	141,055.69
	document or report−related app you use	1,632,897.53
	percentage point of your time that you spend within .5 kilometers of the busara office	3,431,518.01

Heterogeneity by Person ($\tilde{\gamma}_i^{prelim}$)

$$\tilde{\gamma}_i^{prelim} = e^{-\omega^{prelim} z_i} + \tilde{v}_i^{prelim}$$

Low tech skills	1.00
High tech skills	1.10



▲ : included in incentivized naive LASSO decision rule, • : included in incentivized strategy-robust (SR) decision rule. Parameters estimated using GMM as shown, in 2020 week 1 (see Table S6 and Section S2.7.2 to see how it differs from the final model). In cost matrix, off-diagonal elements are regularized to zero ($\Lambda_{offdiagonal} \to \infty$), diagonal elements regularized with $\Lambda_{diagonal} = 1.0$, set via cross validation. $\tilde{v}_i$ plot omits top 5 percent of observations.

Table S5: Performance of Decision Rules (Preliminary vs. Final Cost Estimates)

	Costs		**All Outcomes**			**Income**		
	Preliminary	Final		(pooled)				
	c^{prelim}	c_{kk}	$\boldsymbol{\beta}^{LASSO}$	$\boldsymbol{\beta}^{SR}$	$\boldsymbol{\beta}^{SR_{final}}$	$\boldsymbol{\beta}^{LASSO}$	$\boldsymbol{\beta}^{SR}$	$\boldsymbol{\beta}^{SR_{final}}$
	¢/action^2			.			¢/action	
Panel A: **Decision Rule**								
text_count_out	0.035	0.035	.	.	.	-0.395	-0.107	-0.101
text_count_incoming	0.038	0.037	.	.	.	0.065		
text_count_evening	0.058	0.057	.	.	.		-0.121	-0.120
call_count_out	0.591	0.480	.	.	.	0.625	0.542	0.531
Intercept						301.071	304.622	302.472
λ						1087.036	1087.036	759.296
Panel B: **Prediction Error**								
Baseline Data:				RMSE ($)			RMSE ($)	
Control			3.659 (0.019)	3.732 (0.018)	3.572 (0.019)	3.574 (0.049)	3.583 (0.050)	3.584 (0.059)
Predicted Transparent (prelim costs)			5.830 (0.339)	3.812 (0.028)	3.582 (0.019)	3.620 (0.050)	3.586 (0.054)	3.586 (0.058)
Predicted Transparent (final costs)			7.805 (0.216)	4.025 (0.026)	3.603 (0.017)	3.702 (0.058)	3.591 (0.061)	3.591 (0.058)
Implemented:								
Opaque			3.780 (0.089)	3.710 (0.061)	.	3.549 (0.238)	3.525 (0.252)	.
Transparent			4.641 (0.155)	4.130 (0.096)	.	3.675 (0.219)	3.484 (0.218)	.
N (Control individuals)			1391	1391	.	1376	1376	.
N (Treatment person-weeks, Opaque)			1344	1344	.	75	75	.
N (Treatment person-weeks, Transparent)			1246	1298	.	90	74	.

Notes: The first panel reports the decision rule associated with the challenge, and the costs associated with manipulating these behaviors. The below panels report the performance of each decision rule by outcome, root mean squared error (RMSE) at the week-decision-rule level. Pooled metrics present the mean RMSE across decision rules. Predicted Transparent represents the average expected performance of decision rules given the theoretical model, behavior incentives, and estimated costs. Implemented Transparent/Opaque represents the average performance of decision rules when assigned with/without transparency hints. Lambda regularization differs slightly between model generations due as we altered the protocol to select penalization closer to the boundary of 3 coefficients. SR decision rule estimated using preliminary costs estimates, while SR_{final} is estimated using final costs estimates. RMSE for "Control" and "Predicted Transparent" conditions for pooled rules based on randomly stratified sample across assigned rules: for each outcome, we randomly stratify the set of individuals into equal groups across assigned decision-rule-weeks, and compute the simulated RMSE based on this stratified sample. We then average this RMSE across outcomes to compute the pooled RMSE, and bootstrap this entire procedure with 50 draws to compute bootstrapped standard errors for these pooled RMSE stats.

Table S6: Method Versions

Week	Data		Mistakes	Cost Model		
	Server Issues	Phase 1 Sample	$\hat{\gamma}_i$ Error	z_i	ϕ	Moments
Implemented Decision Rules						
2019-46	Upload Difficulties	Omit last 2 weeks	Yes	Adv. Tech Skills	0.005	Preliminary
2019-47	Upload Difficulties	Omit last 2 weeks	Yes	Adv. Tech Skills	0.005	Preliminary
2019-48	-	Omit last 2 weeks	Yes	Adv. Tech Skills	0.005	Preliminary
2019-49	-	Omit last 2 weeks	Yes	Primary Educ. or Lower	0.005	Preliminary
Break for holidays						
2020-01	-	All weeks	Yes	Adv. Tech Skills	0.005	Final
2020-02	-	All weeks	Yes	Adv. Tech Skills	0.010	Final
2020-03	Blackout	All weeks	Yes	Adv. Tech Skills	0.005	Final
2020-04	-	All weeks	No	Adv. Tech Skills	0.005	Final
Final Decision Rules						
Final †	-	All weeks	No	Adv. Tech Skills	10^{-6} (equivalent to 0.005*)	Final

Notes:: Model implementation issues and version differences week by week. 'Server issues' describes difficulties on the server-side related to data upload and data collection. z_i describes the cost heterogeneity dimension. ϕ describes the shrinkage parameter for unobserved cost heterogeneity. *: note that given the other final model updates, the scale of ϕ changed. $\phi = 10^{-6}$ in the final model produces decision rules that minimize the L2 distance with implemented decision rule coefficients for our final week models of income and intelligence (based on Ravens). †: final method included a few other changes. Further details on these differences described in the text of S2.2.1.

Table S7: Manipulation Can Harm Prediction (Monte Carlo): Cumulative Iterative Approach

	Decision Rule				**Performance** (squared loss)	
	β_1	β_2	β_3	α	On training data	When implemented
Panel A: **Data Generating Process (Unmanipulated)**						
$\boldsymbol{b}^{DGP}$	3.00	0.10	0.10	0.20	0.25	4147.95
Panel B: **Status Quo Approaches**						
$\boldsymbol{\beta}^{OLS}$	3.01	0.10	0.10	0.20	0.25	4198.12
Iterative retraining - Estimated cumulatively						
$\boldsymbol{\beta}^{OLS(2)}$	-0.02	2.84	-2.34	0.16	1.90	1491.03
$\boldsymbol{\beta}^{OLS(3)}$	0.02	0.40	0.79	-0.32	8.56	10.11
$\boldsymbol{\beta}^{OLS(4)}$	0.04	0.35	0.56	-0.72	8.95	8.55
$\vdots$						
$\boldsymbol{\beta}^{OLS(1001)}$	0.36	0.60	-0.05	-1.45	7.98	7.19
$\boldsymbol{\beta}^{OLS(1002)}$	0.36	0.60	-0.05	-1.45	7.98	7.19
Panel C: **Strategy-Robust Method**						
$\boldsymbol{\beta}^{SR}$	0.29	0.50	-0.01	-0.94	7.27	6.86

Notes: Monte Carlo simulation results. Panel A shows the coefficients that relate the outcome to natural behaviors under the data generating process (DGP), $y_i = \mathbf{b}'\underline{\mathbf{x}} + e_i$. Panel B shows coefficients from OLS. For the retraining approach, the training data for $\boldsymbol{\beta}^{OLS(n)}$ is the cumulative manipulated data from all prior periods when $\boldsymbol{\beta}^{OLS(n-1)}...\boldsymbol{\beta}^{OLS}$ were assigned. Panel C shows coefficients estimated with the strategy-robust method. Performance is assessed on the same sample of individuals under the training data, and when the data is manipulated.

Parameters: N=10000, actual costs are $C_i = \frac{1}{\gamma_i}C$, $\gamma_i = 10^{Bernoulli(0.5)}$; policymaker knows C and the distribution of γ_i but not the value for each i; $\tilde{C}_i = \frac{1}{\tilde{\gamma}_i}C$, $\tilde{\gamma}_i = 10^{Bernoulli(0.5)}$. Policymaker beliefs use B=10 draws of $\tilde{\gamma}_i$, see equation (6). $\underline{\mathrm{x}} \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} 1.0 & 1.0 & 0.1 \\ 1.0 & 2.0 & 1.0 \\ 0.1 & 1.0 & 1.0 \end{bmatrix}\right)$, $C = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 2.0 & 0 \\ 0 & 0 & 4.0 \end{bmatrix}$, $e_i \overset{iid}{\sim} N(0, 0.25)$.

Table S8: Manipulation Can Improve Prediction (Monte Carlo)

Decision Rule				**Performance** (squared loss)	
	β_1	β_2	α	On training data	When implemented
Panel A: **If desired targets find it easier to manipulate** ($\zeta = 1$: $cor(\gamma_i, y_i) > 0$)					
$\boldsymbol{b}^{DGP}$	2.00	1.00	1.00	24.62	21.31
$\boldsymbol{\beta}^{OLS}$	1.96	1.02	1.04	24.62	20.47
$\boldsymbol{\beta}^{SR}$	1.88	1.06	-1.88	33.13	10.42
Panel B: **If desired targets find it harder to manipulate** ($\zeta = -1$: $cor(\gamma_i, y_i) < 0$)					
$\boldsymbol{b}^{DGP}$	2.00	1.00	1.00	24.62	83.58
$\boldsymbol{\beta}^{OLS}$	1.96	1.02	1.04	24.62	80.12
$\boldsymbol{\beta}^{SR}$	0.31	1.82	0.95	28.73	29.48

Notes: Monte Carlo simulation results for $y_i = a + \boldsymbol{b}^{DGP}\underline{\mathbf{x}}_i + \underbrace{z_i + \nu_i}_{e_i}$. $\boldsymbol{\beta}^{OLS}$ are estimated coefficients from OLS; $\boldsymbol{\beta}^{SR}$ are coefficients estimated with the strategy robust method, with costs known during training ($\tilde{C}_i \equiv C_i$). Performance is assessed on the same sample of individuals, on training data, or when implemented (with manipulation). When desired targets find it easier to manipulate (Panel A), manipulation improves the performance of all decision rules by signaling part of the unobserved error. In contrast, when desired targets find it harder to manipulate (Panel B), manipulation reduces the performance of all decision rules, as types become muddled. The standard approach does not account for the correlation between ease of manipulation and the target outcome (ζ): it returns the same decision rule regardless of this correlation. In contrast, the strategy robust method can better exploit the signaling value in manipulation when it is aware of the source of heterogeneity. Parameters: N=10000, $\underline{\mathbf{x}}_i \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right)$, $C_i = \frac{1}{\gamma_i}\begin{bmatrix} 2 & 0 \\ 0 & 10000 \end{bmatrix}$, $\gamma_i = e^{\frac{1}{5}\zeta z_i}$, $\tilde{C}_i = C_i$, $z_i \overset{iid}{\sim} N(0, 25)$, $\nu_i \overset{iid}{\sim} N(0, 0.5)$.

Table S9: LASSO Decision Rules Restricted Based on Estimated Costs

	Costs (Actual)	**Income**				**Intelligence** (above median Ravens)			
		$\boldsymbol{\beta}^{LASSO_{final}}$	$\beta_{c_{kk}^{Q50}}^{LASSO_{final}}$	$\beta_{c_{kk}^{Q75}}^{LASSO_{final}}$	$\boldsymbol{\beta}^{SR_{final}}$	$\boldsymbol{\beta}^{LASSO_{final}}$	$\beta_{c_{kk}^{Q50}}^{LASSO_{final}}$	$\beta_{c_{kk}^{Q75}}^{LASSO_{final}}$	$\boldsymbol{\beta}^{SR_{final}}$
Panel A: **Decision Rule**									
text_count_out	0.035	-0.499			-0.101				
text_count_incoming	0.037	0.141				0.270			0.115
text_count_out_evening	0.054				-0.120				
call_count_out	0.480	0.657			0.531				
call_count_outgoing_missed	1.913					-0.156			
calls_noncontacts	1.929					-0.547			-0.519
max_daily_texts_incoming	3.471								0.420
std_dev_evening_text_count	92.37		-0.821				0.673		
avg_duration_evening_calls	19761.46		-0.024						
call_count_60sec_or_more	395022.17		1.200	1.051					
Intercept		296.342	310.400	302.746	302.472	489.685	500.614	507.549	487.033
$\lambda^{decision}$		759.296	759.296	759.296	759.296	1032.37	1032.37	1032.37	1032.37
Panel B: **Prediction Error**		RMSE ($)				RMSE ($)			
Predicted Opaque		3.572	3.602	3.615	3.584	4.972	4.991	4.999	4.973
Predicted Transparent		3.876	3.602	3.615	3.591	4.988	4.991	4.999	4.975

Notes: Panel A reports the decision rules derived from naive LASSO, LASSO models that restrict the feature set to only behaviors that are above the 50th quantile of estimated costs (c_{kk}^{Q50}), the 75th quantile of estimated costs (c_{kk}^{Q75}), and our strategy-robust model. Panel B reports predicted performance, using our experimentally estimated model of manipulation. Predicted Opaque represents the average performance of decisoin rules in control person-weeks, when no behavior was incentivized. Predicted Transparent represents the average expected performance of decision rules given the theoretical model, behavior incentives, and estimated costs. $\lambda^{decision}$ is held fixed at the level that selects three behaviors when allowed to select from all. $\beta^{LASSO_{final}}$ presented in this table differs slightly from the β^{LASSO} which was implemented. The regularization protocol was updated to select penalization closer to the boundary of 3 coefficients and the sample was changed to coincide with that used for the SR model (it includes only individuals with nonmissing tech skills, dropping approximately 1.5 percent of the sample).