Games and Economic Behavior 148 (2024) 449–486

Contents lists available at ScienceDirect

Games and Economic Behavior

journal homepage: www.elsevier.com/locate/geb

Inference from biased polls ☆

Andy Brownback [a,*], Nathaniel Burke [b], Tristan Gagnon-Bartsch [c]

[a] University of Arkansas Department of Economics, 220 N McIlroy Ave #401, Fayetteville, AR, 72701, United States of America
[b] West Virginia University Economics Department, Reynolds Hall, 83 Beechurst Ave #4006, Morgantown, WV, 26505, United States of America
[c] University of Iowa Department of Economics, S252 Pappajohn Business Building, Iowa City, IA, 52242, United States of America

ARTICLE INFO

JEL classification:
D91
D84
D72

Keywords:
Polling
Social desirability
Inference
Signaling

ABSTRACT

People often attempt to present a positive image by overstating virtuous behaviors when responding to unincentivized "polls." We examine whether others account for this "socially desirable responding" (SDR) when drawing inferences from such unincentivized responses. In an experiment, we incentivize "predictors" to guess others' choice behaviors across actions with varying social desirability. Predictors observe random subsamples of either (i) incentivized choices or (ii) hypothetical claims. The hypothetical claims exhibit systematic SDR and predictors are reasonably skeptical of them. However, their skepticism is not tailored to the direction or magnitude of SDR. This under-correction occurs even though subjects' stated sentiment toward the actions can predict SDR.

1. Introduction

Presenting a positive image is a widespread human desire, and many are willing to incur significant costs to do so (Veblen, 1899; Bagwell and Bernheim, 1996; Bursztyn et al., 2018). We therefore expect people to take advantage of opportunities to costlessly inflate their own image. Indeed, in cases where social virtues and stigmas are well-known, people often misreport their views, traits, or behaviors in response to unincentivized elicitations. Such misreporting is known as *socially desirable responding (SDR)* (Maccoby and Maccoby, 1954; Edwards, 1957; Paulhus, 1984), and it arises in a range of settings, including opinion surveys, self-reports, political polls, or simply conversations among friends.

These types of unincentivized elicitations—*polls*, hereafter—are often the best available source of information even when they are plagued by SDR. For instance, doctors rely on self-reports to design treatments in stigmatized domains such as alcohol use or mental health even though these reports exhibit well-documented biases (Del Boca and Noll, 2000; Latkin et al., 2017; Bharadwaj et al., 2017); businesses use political polls to anticipate changes in government policies despite potential bias in such polls (Finkel et al., 1991); and job-seekers rely on information from other workers that may be overly-optimistic about job prospects (Arnold et al., 1985). At the same time, a growing literature finds that many people still respond to polls truthfully, and that simple, unincentivized

☆ For helpful comments, we thank Chiara Aina, Benjamin Bushong, Jonathan de Quidt, Uri Gneezy, David Huffman, Alex Imas, Michael Kuhn, Sherry Li, Peter McGee, Matthew Rabin, Joshua Schwartzstein, Marta Serra-Garcia, and seminar audiences at the SEA Annual Meeting, AEA Mentoring Pipeline Conference, ESA North American Meetings, University of East Anglia's Workshop on Behavioral Game Theory, West Virginia University, University of South Carolina, Virginia Tech University, and La Universidad del CEMA. AEA RCT registry number: AEARCTR-0005186 (available at https://doi.org/10.1257/rct.5186-1.0). We gratefully acknowledge support from the AEA Mentoring Program (NSF Award #1730651).

* Corresponding author.
E-mail addresses: abrownback@walton.uark.edu (A. Brownback), nathaniel.burke@mail.wvu.edu (N. Burke), tgagnonbartsch@uiowa.edu (T. Gagnon-Bartsch).
URLs: https://www.andybrownback.com (A. Brownback), https://www.nathanielcburke.com/ (N. Burke), https://gagnon-bartsch.com/ (T. Gagnon-Bartsch).

https://doi.org/10.1016/j.geb.2024.10.007
Received 1 July 2022

Available online 7 November 2024

elicitations can be useful for predicting behavior (Dohmen et al., 2011).[1] Thus, even though poll data may be biased by SDR, a careful observer of this data may be able to extract valuable information from it.

We experimentally study whether people can anticipate when SDR will (and will not) appear in poll data and if they can then account for it when drawing inferences about choice behavior. Extracting accurate signals from potentially biased poll data requires an appreciation that responses cannot always be taken at face value and an understanding of how they might be distorted. One must anticipate SDR and discount claims of virtuous behavior while also recognizing that people are unlikely to be lying when they admit to stigmatized behaviors. We refer to this ability to interpret poll data in a way that corrects for SDR as *social sophistication.*

In our study, we elicit both poll responses and actual choice behaviors, allowing us to clearly measure the SDR in our poll data. We then examine the degree of social sophistication present when people are given this poll data and asked to predict others' actual choice behaviors. Since we directly observe the effect of SDR on the poll data, we can similarly observe how people correct for it in their predictions. We find evidence of social sophistication along fundamental dimensions: people do anticipate the potential for biased poll data and discount the hypothetical claims of others. However, we find no evidence of more complex dimensions of sophistication: people make costly errors by not tailoring their discounting to the direction or magnitude of SDR.

Our design develops a novel methodology of information provision to identify how beliefs respond to poll data.[2] We reveal random subsamples from an assigned information source (poll data or actual choice behaviors), inducing mechanically random sampling variation in signals. By controlling for differences in the distribution of signals, we isolate signal variation from sampling noise, which reveals experimentally-random changes in signals and their *causal* impact on inferences. Our random assignment of information sources also provides causal evidence on the heterogeneous interpretation of information from different sources.

Specifically, we begin by constructing a setting where we can observe how SDR affects responses to eight separate actions. Six actions involve deciding whether to donate $1 to an organization: St. Jude Children's Hospital, a local NPR affiliate, the Democratic National Committee, the Republican National Committee, Joe Biden's campaign, and Donald Trump's campaign. The other two actions are stealing $1 from another participant in the study and taking $1 for yourself from a planned donation to the Make-A-Wish Foundation.

We measure actual choices and hypothetical claims using parallel elicitations with two distinct groups of subjects. Participants in these groups answer whether they would take each of the above actions. In the *incentive-compatible (IC) group*, we use an incentivized revealed-preference elicitation to measure actual choices. In the *hypothetical (H) group*, we use an unincentivized stated-preference elicitation to measure hypothetical claims about behavior. SDR prompts the H group to overstate (understate) their demand relative to the IC group for actions they believe to be virtuous (stigmatized).

The actions we consider vary in social desirability, but we take no ex-ante stance on which actions tend to be viewed as virtuous. Instead, we recruit a separate *sentiment group* to rate the social desirability of each action. We use this independent evaluation to establish that SDR is well-predicted by sentiment in our controlled setting. A one standard-deviation (SD) increase in how the sentiment group scores an action's social desirability is associated with a 3.1 percentage-point increase in the H group's overstatement of demand for that action ($p < 0.001$). For example, our sentiment group evaluated donating to St. Jude as the most virtuous action (2.24 SD above the mean). This is associated with a 13 percentage-point overstatement of claimed desire to donate: 75% *claim* they would donate, but only 62% do.

We then evaluate the degree to which people anticipate and correct for the SDR manifest in the H group's claims. To do so, we incentivize *predictors* to guess the aggregate choice behavior of the IC group for each action.[3] Predictors make initial guesses about choice behavior. They are then randomly assigned to observe "signals," which are subsamples of either (i) choices from the IC group itself or (ii) claims from the H group. Predictors then make updated guesses about the behavior of the IC group. By observing predictors' updating behavior, we can deduce the differential weighting of information from the two sources and evaluate key hypotheses about their social sophistication. Specifically, we assess whether predictors account for SDR by appropriately discounting the claims of the H group.

Our first main hypothesis examines whether predictors anticipate SDR and accordingly down-weight the (potentially biased) claims from the H group. We find that they do. 31% of predictors' updating from IC-group signals is "extra updating" attributable to the added weight given to the IC-group's choices relative to the H-group's claims ($p < 0.01$).[4]

In social interactions, people often have experience both reporting their views to others—similar to our H group—and drawing inferences based on others' reports—similar to our predictors. For this reason, we designed our study to examine how prior experiences may influence social sophistication. To do so, we compare the updating of newly recruited predictors to those who previously participated in either the IC or H group. We find suggestive but inconclusive evidence that predictors who previously participated in the H group discount the claims of the H group more than predictors without this experience ($p = 0.125$). These subjects may be more skeptical of hypothetical claims because they experienced the impulse to lie when making such claims.

[1] People may respond truthfully both because of a preference for being honest and a preference to appear honest, as documented by a large experimental literature on an aversion to lying; see, e.g., Abeler et al. (2019) for a meta-study of 90 studies using designs similar to Fischbacher and Föllmi-Heusi (2013).

[2] A large literature demonstrates that information provision influences beliefs and attitudes across numerous policy-relevant domains; see Haaland et al. (2020) for a review.

[3] Predictors are a mix of newly recruited participants and returners from the IC and H groups. As we detail later, this allows us to examine how experience with SDR affects predictor behavior.

[4] All estimates presented in the introduction are derived from our within-subjects specification. See Section 5 for details on our analysis.

While discounting the average signal from the H group is a fundamental part of sophisticated inference, discounting all signals equally would not reflect full social sophistication. Full sophistication calls for people to adjust their discounting depending on the direction and magnitude of the bias—the focus of our second and third main hypotheses.

Our second main hypothesis explores more complex social sophistication by asking whether predictors recognize the direction of SDR; that is, whether it is socially desirable to overstate or understate demand for an action. Social sophistication rests on such knowledge as it enables a predictor to determine when they should discount an H-group signal because it is likely biased in the direction of SDR and when they should give it additional weight because it runs against the direction of SDR. We define a signal to be "SDR-congruent" if it implies a greater frequency of socially desirable choices *than initially guessed*. A signal is "SDR-dissonant" if it implies a lower frequency *than initially guessed.*[5] An SDR-congruent signal from the H group should be discounted because it likely reflects inflated claims. In contrast, an SDR-dissonant signal from the H group is particularly informative because it suggests that more respondents than expected admit to socially undesirable behavior despite the opportunity to freely claim virtuous behavior. We find that predictors fail to recognize this. While they correctly discount SDR-congruent signals from the H group by 18% relative to the IC group ($p < 0.001$), they treat SDR-dissonant signals from the H group almost identically to those from the IC group.

Our third main hypothesis asks if predictors recognize the relative magnitude of SDR across the eight actions. When considering actions that are notably biased, predictors should treat claims from the H group with increased skepticism. However, we find no evidence that predictors discount signals for each action based on the degree of SDR for that specific action. If anything, our point estimates suggest that predictors' guesses place *more* weight on claims from the H group as SDR becomes more extreme ($p > 0.10$).

The lack of social sophistication demonstrated by predictors' guesses stands in striking contrast to the responses of the sentiment group. When explicitly asked to evaluate the social desirability of each action, the responses of the sentiment group were highly predictive of which actions would exhibit greater SDR. Hence, our population does have knowledge of which actions tend to incite greater social-image concerns. Yet, it appears that predictors neglect this knowledge when deciding how to evaluate claims from the H group.

We also find irregularities in the confidence predictors place in their guesses. After each guess cast by predictors, we elicited their confidence in that guess. For initial guesses, we find a negative correlation between the accuracy of a predictor's guess and their confidence ($p < 0.01$). This "Dunning-Kruger" effect (Kruger and Dunning, 1999) persists for updated guesses among predictors who receive information from the H group ($p < 0.05$), but such false confidence is diminished for predictors in the IC group who receive higher-quality information. In line with the limited social sophistication we find elsewhere, predictors show no differences in average confidence when receiving information from the H or IC group, suggesting they do not realize the superiority of the IC-group information.

One concern in comparing signals from the H and IC groups is that predictors may not treat them as equally reliable. We address two beliefs that may motivate this behavior. First, uncertainty in (or polarization around) an action's sentiment may give rise to more erratic claims from the H group, causing some to overstate claimed demand while others understate claimed demand. Predictors may then have uncertainty about the direction that SDR biases the H group's claims. Second, regardless of whether respondents attempt to respond in a socially-desirable manner, the H-group signals may simply be noisier because of a lack of incentives to focus. We address the first concern by constructing a proxy for the noise inherent in signals about an action—the variance across evaluators in the sentiment scores associated with the action. Disagreement or uncertainty in the sentiment group captures uncertainty about the direction of any biases and also captures the degree of polarization towards an action. Using this proxy variable, we compare the discounting of H-group signals across actions whose signals have varying degrees of reliability. We find no clear associations between sentiment variance and discounting. We also find that this proxy for signal reliability is, in fact, negatively associated with confidence in guesses. The same confidence data can help us address the second concern. If predictors discount H-group signals purely because of concerns about "white noise," then these lower-quality H-group signals should lower their confidence relative to the higher-quality IC-group signals. As we described in the paragraph above, this is not the case. Furthermore, for both concerns, if noise were the primary driver of discounting, then confidence should diminish as noise increases. As mentioned, this does not happen as noise from sentiment uncertainty rises; thus, it is difficult to imagine that it would happen as a result of anticipated "white noise." Predictors' discounting behavior therefore appears inconsistent with concerns about the noise in H-group signals nor the polarization of an action.

The benefits of having accurate data when making economic decisions are clear. Because of the difficulties we find in how people process data affected by SDR, the benefits of accuracy may further increase in the presence of SDR. Fortunately, researchers have developed several tools, such as the randomized-response technique (Warner, 1965) and list experiments (Raghavarao and Federer, 1979; Karlan and Zinman, 2012), to identify underlying preferences when SDR is prevalent and incentivized elicitations are not possible. These tools have identified SDR in a broad set of stigmatized and virtuous domains.[6] Moreover, Rosenfeld et al. (2016) find that these techniques can correct biased estimates and improve inference from polls. In light of our results, we believe there is strong

[5] Note that there are no objectively high or low signals, only higher or lower signals than initial guesses. Under this definition, even an SDR-dissonant signal from the H group is consistent with SDR. SDR suggests that more people hypothetically claim socially desirable behavior than actually choose it, while an SDR-dissonant signal shows fewer people claiming socially desirable behavior relative to *the predictor's initial guess* of how many would actually choose it.

[6] Tourangeau et al. (2000) and Tourangeau and Yan (2007) provide reviews. SDR has been identified in political polls—often called a "Bradley Effect" or "Shy Tory Factor" (Reeves et al., 1997; Hopkins, 2009; Brownback and Novotny, 2018); polls for female and minority candidates (Heerwig and McCabe, 2009; Streb et al., 2008; Stephens-Davidowitz, 2014; Kane et al., 2004; Brown-Iannuzzi et al., 2019); sentiment surrounding race (Krysan, 1998), immigration (Janus, 2010), and same-sex marriage (Powell, 2013; Coffman et al., 2017); revelation of vote-buying behavior (Gonzalez-Ocantos et al., 2012); voter turnout (Holbrook and Krosnick, 2010); and religious attendance (Jones and Elliot, 2016).

evidence in favor of using these tools in regular audits to identify SDR and recalibrate polls. Independent sentiment surveys could also be used to predict susceptibility to SDR.

SDR is typically understood as a means of projecting a positive image of oneself, likely as a combination of both social- and *self*-signaling (e.g. Bénabou and Tirole, 2002). These dual motivations may explain why SDR persists in many online and anonymous contexts such as ours. While this anonymous context likely mutes the impact of SDR, our sentiment-group results show that SDR is still present and predictable. Additionally, it provides a test of our key hypotheses in a relevant context since real-world polls often employ anonymity in an attempt to address SDR and experimenter-demand effects.[7]

Our exploration of social sophistication advances the literature on social norms in general and on SDR specifically. Krupka and Weber (2013) demonstrate that social norms—similar to stigma and virtue in our domain—are well-anticipated by experimental subjects. A conceptually related paper on "political correctness," Braghieri finds that SDR creates a "wedge" between public and private statements, reducing the information content of public statements. Our paper complements this analysis by exploring the wedge between private statements and consequential choices. Subjects in Braghieri are able to anticipate discrepancies between public and private statements, but similar to our findings, subjects exhibit limited sophistication when predicting heterogeneity in the bias. Design differences between the two studies may be informative about the mechanisms at play. Braghieri explicitly asks subjects to predict discrepancies between private and public statements, while we use a difference-in-differences design to indirectly elicit perceived discrepancies between information sources. An explicit elicitation may prompt subjects to consider the possibility of misreporting and hence may explain why subjects in Braghieri exhibit greater sophistication than subjects in our study.[8]

Our study also relates to a broader literature on information extraction from potentially biased communication. Crawford and Sobel (1982) develop the notion of "cheap-talk" equilibria and show how receivers can extract information from signals even when senders have misaligned incentives. In a setting with similarly misaligned incentives, Kartik (2009) demonstrates the informativeness of communication when senders bear some cost of misreporting their private information. Although experimental studies demonstrate the benefits of communication even when incentives are not aligned (see Farrell and Rabin, 1996 and Crawford, 1998 for reviews), our results suggest that these benefits may be limited.

There are also connections between our study and a growing experimental literature on failures to extract others' private information in strategic contexts due to the difficulties of hypothetical and contingent reasoning (e.g., Esponda and Vespa, 2014; Martínez-Marquina et al., 2019; Araujo et al., 2021; Ngangoué and Weizsäcker, 2021). For instance, Esponda and Vespa (2014) find that many subjects in a simultaneous voting problem fail to condition on the hypothetical event of being pivotal and consequently fail to properly extract information. In our setting, predictors may similarly fail to contemplate how H-group choices would have differed if they had been incentivized. Esponda and Vespa (2014) also show that subjects were able to extract information more successfully in a sequential voting problem where there is no need to think hypothetically. This may relate to our finding that subjects appear knowledgeable about the prevailing sentiment toward our actions when explicitly asked, yet fail to fully leverage this knowledge when drawing inferences from H-group signals.

Our paper begins with an explanation of our experimental design in Section 2. Section 3 follows with our hypotheses and a simple model that develops the intuition behind them. We then evaluate these hypotheses in Sections 4 and 5. Section 6 concludes.

2. Experimental design

Our study design was pre-registered with the AEA RCT registry. It consisted of three parts: the Sentiment Study, the Choice Study, and the Prediction Study. Each part took place online with subjects recruited from the University of Arkansas.

We recruited 39 subjects for the Sentiment Study. For the Choice Study, we recruited 187 subjects and split them into two groups, which we call the IC group and H group (described in detail below). In the Prediction Study, we recruited 95 new subjects to combine with returners from the Choice Study. Table 1 breaks our sample down by assignment.

2.1. Actions

In each part of our study, subjects considered binary choices which involved taking an action or not. The same eight actions where featured throughout the experiment:

St Jude Donation: Donate $1 to the St. Jude Children's Hospital.
NPR Donation: Donate $1 to KUAF radio station, the local NPR affiliate.
Steal: Steal $1 from a participant in another part of the study.
Take Donation: Take $1 for yourself from a planned $50 donation to the Make-A-Wish Foundation.
Trump Donation: Contribute $1 to Donald Trump's presidential campaign.
Biden Donation: Contribute $1 to Joe Biden's presidential campaign.

[7] "Experimenter demand"—where subjects respond in a manner they perceive to be consistent with the experimenter's intention—is one expression of SDR. Although de Quidt et al. (2018) find that the impact of this responding bias may be limited, our results suggest that lay observers of biased experimental data are unlikely to accurately predict the direction or degree of the bias.

[8] Charness et al. (2021) similarly examine how subjects evaluate information from biased sources. However, we ask whether people can identify and correct for biased poll data, while they ask whether subjects can optimally select between biased data sources. They find that subjects tend to over-select sources that provide confirmatory evidence.

Table 1
Subject participation by treatment.

	Sentiment Study	Choice Study	Prediction Study
Sentiment Group	39 Subjects		
IC Group		91 Subjects	84 Returners 7 Non-Returners
H Group		96 Subjects	92 Returners 4 Non-Returners
New Predictors			95 New Subjects
Totals	39 Subjects	187 Subjects	271 Subjects

Notes: "Non-Returners" failed to complete the Prediction Study after successfully completing the Choice Study.

RNC Donation: Contribute $1 to the Republican National Committee.
DNC Donation: Contribute $1 to the Democratic National Committee.

We made no attempt to label the predominant valence of an action (e.g., "virtuous" or "stigmatized") based on our a priori perceptions. We designed our experiment and all hypotheses to be agnostic about the sentiment surrounding actions; instead, we categorize actions empirically based on the behavior of our subjects, who are all drawn from the same population. In this way, all of our tests could be based on perceptions that are observably present in the population. Moreover, since all subjects were recruited from a relatively homogeneous population of college students, we expect sentiment within our experimental setting to be both more consistent and more well-known than in settings involving more general populations. This design choice was meant to provide our subjects with a better chance to demonstrate social sophistication. Thus, our results reflect an upper bound on social sophistication for populations that are more diverse than ours.

Some discussion of our set of actions is warranted. First, though the predominant emotional valence of any specific action was unimportant to our design, we selected actions to induce a wide range of emotional valence so that we could test for sensitivity to *differences* in social desirability. Second, we selected several political actions because SDR is commonly observed in political settings.[9] Thus, measuring SDR and testing for sophistication about it would be natural in this domain. Third, many actions were chosen in pairs so that the sentiment surrounding them would be likely to covary negatively. These steps were taken to increase the variance in choice behaviors and predictions so that we would not spuriously attribute general behaviors to systematic differences in behavior resulting from SDR.

The binary nature of decisions (either to take an action or not) simplified the experiment and allowed us to send easily-understood signals of behavior to our predictors. All choices were made privately through online surveys. Subjects were assured that no individual responses would ever be viewed by anyone except the researchers. This is a conservative approach that likely mutes the impact of social desirability, since SDR is often dependent on the anticipated reactions of observers. As previously discussed, this provides a more natural test of social sophistication about SDR without experimenter demand effects. Actions were described identically and in detail to all subjects in all parts of the study, including information about the anonymity under which choices and statements were made. See Appendix Section C (and Figs. C.1-C.7 therein) for the full description given to subjects.

2.2. Sentiment study

We recruited 39 subjects to evaluate the sentiment associated with each of the eight actions listed above. Subjects who participated in the Sentiment Study did not participate in any other portion of the experiment; they were paid a flat fee of $5.

For each of our eight actions, subjects answered the three questions below on a scale of 0-10, where 0 represented "Very Negative" and 10 represented "Very Positive" sentiment.

1. How would you feel about taking this action yourself?
2. How would you feel about other people who take this action?
3. How do you think most other people would feel about people who take this action?

For each action A, let $Q_{i,j,A}$ denote subject i's response to question $j \in \{1,2,3\}$ above. We then construct subject i's "perceived virtue" of action A, denoted $V_{i,A}$, by taking the within-subject mean of these responses: $V_{i,A} \equiv \frac{\sum_{j=1}^{3} Q_{i,j,A}}{3}$.[10] Letting N_S denote the number of subjects in the Sentiment Study, we will use the following indices to measure the perceived virtue of action A:

[9] See, e.g., Reeves et al., 1997; Streb et al., 2008; Stephens-Davidowitz, 2014; Hopkins, 2009; Brownback and Novotny, 2018.

[10] We asked multiple sentiment-related questions in order to capture first- and higher-order beliefs about social desirability that may influence SDR. Our results hold if we replace the composite measure of sentiment with any of the individual measures; see Sections 4.1 and Appendix Section A.1 for details.

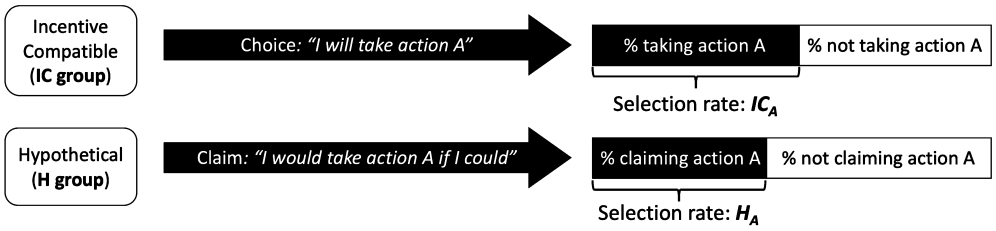

Fig. 1. Experimental Design: Choice Study.

$$V_A \equiv \frac{\sum_{i=1}^{N_S} V_{i,A}}{N_S}, \tag{1}$$

$$\widehat{V}_{i,A} \equiv \frac{V_{i,A} - \overline{V}_i}{\sigma_i}, \tag{2}$$

where $\overline{V}_i$ and σ_i are subject i's mean and standard deviation of $V_{i,A}$ across all eight actions.

Our pre-registered measure of social desirability, V_A, captures the perceived virtue of action A averaged across individuals. This measure suffers from a lack of statistical power since each action has only one observation. To leverage our full sample of sentiment data and increase statistical power, we replicate our analyses using $\widehat{V}_{i,A}$, which normalizes responses within each individual.

2.3. Choice study

In the Choice Study, subjects evaluated all eight actions after being assigned to one of two groups. The first group, the "IC" group, revealed their preferences through choices in an incentive-compatible elicitation. The second group, the "H" group stated their preferences through claims in a hypothetical elicitation. The IC group had 91 subjects and the H group had 96 subjects.[11]

The only difference between the IC and H groups was the incentive-compatibility of the IC group's choice elicitation. For instance, if a subject in the IC group chose to donate \$1 to St. Jude, then that subject actually sacrificed \$1 of their payment and St. Jude actually received a \$1 donation. If a subject in the H group made such a claim, they sacrificed nothing and St. Jude received nothing. Unlike subjects in the IC group, those in the H group faced no explicit incentives to make claims consistent with their true preferences.

For each action A, let $IC_A \in [0\%, 100\%]$ and $H_A \in [0\%, 100\%]$ denote the "selection rate" for action A among the IC and H group, respectively. We then define socially desirable responding (SDR) as the overstatement (or understatement) of demand for an action when subjects did not have to pay the cost of taking the action:

$$SDR_A \equiv H_A - IC_A. \tag{3}$$

We consider action A to be socially desirable if $SDR_A > 0$; that is, the H group inflated their claimed desire to take that action relative to the choices of the IC group. In contrast, action A is socially undesirable if $SDR_A < 0$. It is important to emphasize that these labels of "desirable" or "undesirable" just indicate whether, on average, the share of people who claim to take the action is higher or lower than the true population choice rate, respectively. In this sense, "desirable" should simply be interpreted as "over-reported." As we discuss more in Section 3, an action labeled as socially desirable under this definition need not be viewed as desirable or virtuous by all—or even most—of the population.

Fig. 1 depicts the flow of the Choice Study for an example where the H group understates demand for an action (i.e. $SDR_A < 0$).

All subjects received a \$5 participation payment in the Choice Study. This amount was subject to change for the IC group because one of their decisions was randomly selected to be binding (e.g., if they chose to donate to St. Jude, and this decision was randomly selected to bind, their payment would decrease by \$1 and St. Jude would gain \$1). All subjects in the Choice Study were told that they must participate in an additional part (the Prediction Study, described below) during which they could earn more money; subjects were not given any description of this additional part until the Prediction Study began.

2.4. Prediction study

In order to receive their full payment, all of the subjects who participated in the Choice Study were required to participate as "predictors" in the Prediction Study, which started five days later. In addition, we recruited 95 new predictors who had not participated in any previous part. In total, the Prediction Study featured 271 subjects: 84 returners from the IC group, 92 returners from the H group, and 95 new predictors. All subjects received a \$5 participation payment for completing this part along with any earnings gained from accurate predictions.

[11] We restricted subjects to participate in the Choice Study only once—either in the IC or H group. We dropped the second submission of anyone who violated this restriction. As a result, we dropped 15 submissions from the IC group and 7 from the H group. While our recruiting system ensured that each registered email was only invited to one group, students who used multiple emails could register in the system twice. Duplicates were identified when we requested on-campus emails from every participant.

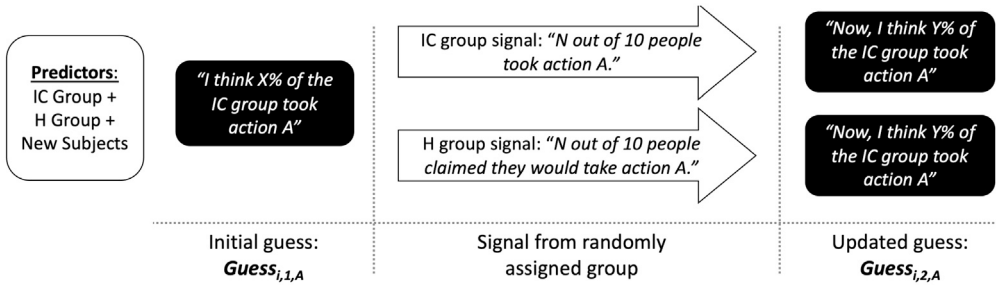

Fig. 2. Experimental Design: Prediction Study.

In the Prediction Study, predictors observed the exact same descriptions of the actions as subjects in the Choice Study and were tasked with guessing the choice behavior of the IC group for each of the actions.[12] To simplify the procedure, we asked subjects to guess what share of the IC group (between 0 and 100, inclusive) chose to take each action. We incentivized predictions using a Becker-DeGroot-Marschak mechanism (Becker et al., 1964).[13]

For each of the eight actions, predictors made two guesses about the IC group's selection rate, IC_A, one before receiving information and one after. Let $\text{GUESS}_{i,1,A}$ denote predictor i's initial guess. Each predictor was then given a randomly drawn "signal" revealing selections from either the IC or H group. Rather than observing the full selection rate, predictors observed a random sub-sampling of behavior. Specifically, predictor i received a signal, $s_{i,A} \in \{0, 1, \ldots, 10\}$, conveying the selections on action A of 10 randomly-sampled respondents from their assigned group.[14] Thus, for information from the IC group, $s_{i,A} \sim \text{Bin}(10, IC_A)$; for information from the H group, $s_{i,A} \sim \text{Bin}(10, H_A)$. Note that these signals were drawn with two independent sources of randomness that are critical to our novel identification strategy: random assignment of the information source—the IC or H group—and random sampling of the information *conditional on its source*.

We gave predictors detailed information about the choice procedures of their assigned group so that they could appropriately tailor the weight given to these signals. We then required predictors to complete a comprehension quiz on the procedures before advancing.[15]

After receiving signals, each predictor submitted updated guesses about the selection rate. Let $\text{GUESS}_{i,2,A}$ denote predictor i's updated guess about the selection rate, IC_A. Fig. 2 depicts the flow of the Prediction Study.

Immediately after revealing their guesses, predictors stated their confidence in each of their guesses. This confidence was elicited on a scale from 0 (very uncertain) to 10 (very confident). This elicitation was not incentivized.

3. Primary hypotheses

Our research questions focus on our notion of "social sophistication." We define social sophistication as actively anticipating SDR and appropriately weighting claims from the H group based on their susceptibility to SDR. To assess the extent to which predictors account for SDR, we measure the weight they assign to (potentially biased) signals from the H group relative to the weight assigned to signals from the IC group. Social sophistication requires that predictors both (i) anticipate the existence of SDR and (ii) adjust for the direction and magnitude of the bias.

To develop intuition for social sophistication, we present a stylized model of SDR and derive hypotheses regarding how a socially-sophisticated Bayesian would respond to information that is subject to SDR. We address the relative weight that should be given to responses that may be biased by SDR and how this weighting depends on perceptions of the direction and magnitude of SDR. We then evaluate these hypotheses in Sections 4 and 5.

Recall that the H group faces no explicit incentives based on their claims. Hence, it is costless for them to *claim* they would take a socially desirable action if given the opportunity. In contrast, the IC group must face the consequences of their choices. For simplicity, we refer to choices made with consequences as revealing "true" preferences.

[12] As mentioned in Subsection 2.3 (Footnote 11), some subjects violated the restriction for duplicate participation. We discovered these duplicates after the Prediction Study, meaning that signals about the IC group were drawn prior to dropping these duplicates. Accordingly, predictors were incentivized based on responses from the full dataset. Choice rates with and without duplicate participants never differ by more than 1.3 percentage points per action. We limit our analysis to non-duplicate predictors in order to honor our experimental protocols. However, our manipulation checks in Tables 3 and 4 use the full dataset, because that is the dataset from which signals were drawn and guesses were incentivized.

[13] Predictors stood to gain an extra $5 payment based on the outcome of a lottery. The probability of winning the lottery was either (a) a random draw from a uniform distribution from 0 to 1, or (b) equal to IC_A. Predictors were paid based on option (a) unless their prediction of IC_A exceeded their random draw from option (a); in this case, they were paid based on option (b).

[14] More specifically, $s_{i,A}$ counts the number of 10 randomly-chosen respondents who elected to take action A. A predictor received such a signal for each action, and thus received 8 signals in total. A predictor received all signals from the same group: either 8 signals from the IC group or 8 from the H group. For each $s_{i,A}$, we randomly drew 10 respondents with replacement.

[15] See Appendix Section C for the exact instructions and comprehension questions.

Suppose that, due to the lack of consequences, there exists a fraction $\theta_A \in [0,1]$ of subjects in the H group who claim a preference toward action A in the way they find socially desirable regardless of their true preference.[16] If participants universally view action A as virtuous, then such a bias leads subjects in the H group to inflate their claimed desire to take the action relative to the IC group. The expected selection rate in the H group is then $H_A = (1-\theta_A)IC_A + \theta_A$: a fraction $1-\theta_A$ of subjects reveals their true preference, and the remaining fraction, θ_A, claim they would take action A regardless of their true preference. Our measure of SDR for action A (Equation (3)) is therefore $SDR_A = H_A - IC_A = \theta_A(1-IC_A)$.[17]

If instead, participants view that A as stigmatized, then H-group subjects will deflate their claimed desire to take the action. Their expected selection rate is then $H_A = (1-\theta_A)IC_A$: a fraction $1-\theta_A$ of subjects again reveals their true preference, while the remaining subjects claim they would refuse the action. Our measure of SDR in this case is $SDR_A = -\theta_A IC_A$.

Recall that for each action A, a predictor in our experiment observes the choices of 10 random subjects from either the IC or H group. Thus, IC-group signals are draws from $\text{Bin}(10, IC_A)$, while H-group signals are draws from $\text{Bin}(10, H_A)$. In our simple framework, H_A clearly depends on θ_A, but we do not assume perfect knowledge of θ_A in developing our hypotheses about social sophistication (nor do we test for such perfect knowledge in our analysis). Our hypotheses hold under uncertainty about the precise value for θ_A and focus on directional predictions about how knowledge of θ_A influences the relative weight given to signals from the IC and H groups.[18]

Before turning to our hypotheses, it is worth noting that people may naturally vary in whether they find a given action virtuous or stigmatized. The simple model above assumes a common understanding of the socially-desirable response across the population. It is straightforward to allow for heterogeneity. More generally, we can interpret θ_A as the fraction of the population who is willing to misreport their behavior when in the H group. Suppose that among those willing to misreport, a faction $\gamma_A \in [0,1]$ views A favorably. The expected selection rate in the H group is then $H_A = (1-\theta_A)IC_A + \theta_A\gamma_A$. When $\gamma_A = 1$, we return to the case of a unanimously virtuous action discussed above; similarly, $\gamma_A = 0$ corresponds to a unanimously stigmatized action.

Allowing for the sentiment surrounding an action to vary across individuals requires us to clarify our language regarding "virtue" and "stigma." Recall that $SDR_A > 0$ is equivalent to action A being *over-reported* (in expectation) by the H group relative to the IC group, while $SDR_A < 0$ is equivalent to A being *under-reported*. We sometimes abuse terminology by using "virtuous" or "stigmatized" as synonymous with over-reported and under-reported, respectively. This terminology does not necessarily mean that A is universally viewed as virtuous or stigmatized. Rather, we take "virtuous" to simply indicate that, among those who misreport their preferences, the majority do so by claiming to take action A, and we use "stigmatized" analogously. These definitions give no consideration of the prevalence of truthful reporting in either direction. To give an example with heterogeneous sentiment, suppose that 90% of the population loathes action A and they neither take nor claim to take that action. The other 10% view A positively and claim to take action A, yet only half of this 10% actually takes action A when facing incentives. Although most of the population views A negatively, it is nevertheless over-reported.[19] We will empirically examine how heterogeneous sentiment may influence our results in Section 5.5.

3.1. Socially desirable responding

We begin our analysis with manipulation checks to demonstrate (i) SDR is present and predictable—overstatement of claimed demand from the H group correlates with evaluation of virtue from the sentiment group—and (ii) predictors place positive weight on signals—the quality of predictors guesses correlates with the quality of their signals.

Confirming these manipulation checks ensures that claims from the H group do, in fact, possess information relevant for predicting the choices from the IC group. These manipulation checks also rule out the possibility that all differences in the two information sources can be wholly attributed to beliefs about noise or random choice errors. If this were the case, then social sophistication would not provide any improvement toward a predictor's guesses.

Manipulation check 1 *(SDR). Socially desirable responding will cause the H group to overstate their claimed demand for an action by more when its perceived virtue grows.*

All else equal, the more virtuous an action is perceived to be, the more beneficial it is to portray oneself as a type who takes that action. Thus, an increase in perceived virtue should increase the incentive to overstate claimed demand. That is, more subjects are inclined to lie in the socially desirable way when incentives are removed.[20]

[16] Equivalently, a fraction $1-\theta_A$ of subjects in the H group report honestly despite no explicit incentive to do so. This could be driven, for instance, by a preference for truth-telling (e.g., Abeler et al., 2019).

[17] This formulation assumes that if a person's hypothetical behavior deviates from their incentivized behavior, then it does so in the direction they find desirable. However, deviations may also occur due to noise. Moreover, people may have varying opinions on which direction is desirable. Such noise or heterogeneity can be incorporated into our approach, and we discuss the interpretation of our results in light of these factors in Section 5.5.

[18] Although signals may influence a predictor's estimate of θ_A, we require them to also influence beliefs about IC_A.

[19] While this is a concern with our modeling approach, it is not one in practice. In Section 4, we show that over-reporting tends to occur in actions that are evaluated positively.

[20] This can be thought of as a local phenomenon, given that it implicitly holds choice rates, IC_A, fixed. This is one limitation of such a stylized model, because a substantial change in the virtue or stigma of an action may affect choice rates. However, this should have little impact on our results, since our actions are all within a reasonable range of stigma or virtue and, empirically, we find that the degree of over-reporting across our actions is uncorrelated with their choice rates.

Confirming the predictive validity of our measures of perceived virtue from the sentiment group establishes what we call "sentiment sophistication." That is, the evaluations of sentiment we collect represent useful social knowledge for predicting choice behavior.

Our second manipulation check evaluates whether predictors respond to signals.

Manipulation check 2 *(Accuracy). If predictors assign positive weight to their signals, then they will be relatively more accurate with information from the IC group.*

Our design cannot identify social sophistication if predictors never update their guesses in response to signals. We present a simple test to demonstrate that predictors assign positive weight to signals—we measure if updated guesses about the IC group are more accurate for predictors who receive their signals from the IC group rather than the H group. That is, do more accurate signals result in more accurate guesses?

3.2. Social sophistication

We proceed by evaluating hypotheses about social sophistication—the anticipation and correction for SDR. For these hypotheses, we use our stylized framework to describe how predictors should respond to signals when they have a sophisticated understanding of how H-group signals differ from IC-group signals.

In a comprehensive review, Benjamin (2019) describes how prevalent biases in statistical reasoning—independent of the concepts we study here—can generate both over- and under-updating from new information. For this reason, all of our hypotheses about updating focus on how updating differs in response to IC-group and H-group signals rather than how updating compares to the Bayesian benchmark. In this way, we can evaluate social sophistication in isolation instead of evaluating the joint test of social sophistication *and* statistical sophistication.[21] Moreover, since we do not restrict a predictor's prior beliefs, the Bayesian benchmark is not readily derived. We would need to elicit each subject's full prior probability distribution over behavior in the IC group in order to derive the Bayesian posterior conditional on their signal.[22]

Our first hypothesis is about anticipating SDR. Social sophistication allows predictors to leverage signals from the H group to make unbiased guesses about IC_A. However, those guesses will be inherently noisier. Sophisticated predictors will recognize that signals from the H group carry less information and will discount them relative to the more informative signals from the IC group. Thus, with social sophistication, updated guesses about the behavior of the IC group will react more strongly to signals from the IC group.

Hypothesis 1 *(Anticipation of SDR).* Predictors with social sophistication will give greater weight to incentive-compatible information.

Hypothesis 1 tests a fundamental aspect of social sophistication. In our stylized model, testing Hypothesis 1 simply amounts to testing whether predictors treat θ_A as non-zero.

Our second hypothesis concerns the direction of SDR. We are interested in the differential weight that predictors attach to H signals that are aligned with the direction of SDR versus those that push against it. To state our hypothesis, we must first designate which signals are aligned with SDR. Let $s^*_{i,A}$ be predictor i's cutoff such that his updated guess would move upward if he were to receive an IC-group signal above that cutoff (i.e. $s_{i,A} > s^*_{i,A}$) and downward if he were to receive an IC-groups signal beneath that cutoff (i.e. $s_{i,A} < s^*_{i,A}$). For instance, $s_{i,A} > s^*_{i,A}$ suggests that more people take action A than he originally guessed. How should the predictor respond to such a signal when it instead originates from the H group? Signals from the H group must be weighted differently, and these weights importantly depend on whether action A tends to be over-reported or under-reported. First suppose it is over-reported (i.e., $SDR_A > 0$). In this case, we say an H signal above $s^*_{i,A}$ is *SDR-congruent*—it is suggestive of more people acting in the socially desirable way. Such a signal is likely the result of inflated claims and thus should be discounted relative to an IC signal of the same magnitude. In contrast, suppose A is under-reported (i.e., $SDR_A < 0$). We then say an H signal above $s^*_{i,A}$ is *SDR-dissonant*—it is suggestive of more people acting in the *undesirable* way. These relatively rare signals from the H group should be given *additional* weight relative to an IC signal of the same magnitude. This is because they likely understate the true rate that A is chosen among the IC group.[23] Analogous to the previous definitions, we say a signal from the H group below $s^*_{i,A}$ is SDR-dissonant when A is over-reported and SDR-congruent when A is under-reported. Since we do not observe $s^*_{i,A}$, in practice, we assume a predictor's initial guess serves as this cutoff.

[21] Our focus on differential updating across groups also mitigates concerns about anchoring. Since we elicit each subject's initial and updated guesses, they may update insufficiently if the latter is anchored toward the former. However, by focusing on differential updating across groups, we largely sidestep this issue.

[22] Appendix A.6 provides some simulated examples that depict a predictor's updated guesses under the Bayesian benchmark for different specifications of the predictor's prior. These plots highlight how the optimal guesses should differ when receiving H-group signals versus IC-group signals.

[23] SDR predicts that SDR-congruent signals will be more likely from the H group than the IC group. Indeed, SDR-congruent signals are 16 percentage points more likely from the H group ($p < 0.001$). For this reason, Hypothesis 1 suggested that the claims of the H group should be discounted relative to the choices of the IC group, *on average*.

Hypothesis 2 *(Direction)*. Predictors with social sophistication will discount "SDR-congruent" signals from the H group relative to signals with the same magnitude from the IC group, but they will give greater relative weight to "SDR-dissonant" signals from the H group.[24]

For example, consider donations to St. Jude, which tend to be over-reported ($SDR_{St.Jude} > 0$). Suppose a predictor initially guesses that 50% of the IC group would donate and then receives a signal in which 60% of the sampled members of the H group claim they would make the donation. This signal is SDR-congruent. It should be discounted relative to a signal in which 60% of the sampled members of the IC group actually choose to donate because the claims from the H group are likely overstated. But, if the same predictor with the same initial guess instead receives a signal from the H group in which only 40% claim they would make the donation, then this signal is SDR-dissonant. It should be treated as *even more* informative than a signal from the IC group in which 40% choose to donate: if only 40% claim they would donate despite being able to freely lie, then surely the true choice rate is even lower than that.[25]

We have described this hypothesis in terms of individual predictors, but this introduces endogeneity into our statistical tests by categorizing signals based on the predictor's initial guess. For our hypothesis testing, we rely on the same intuition but avoid this endogeneity concern by categorizing signals based on population-average initial guesses. See Section 5.2 for details on our model specification.

To summarize, evaluating this hypothesis will jointly test if predictors (i) identify whether SDR inflates or deflates hypothetical claims about a given action A and (ii) understand that this makes SDR-dissonant signals from the H group less likely and, therefore, more informative about true choice rates.

Our third hypothesis relates to the magnitude of SDR. The information content of a signal from the H group is decreasing in the share of subjects falsely claiming socially desirable behaviors, θ_A. Consider, for example, a case where action A is commonly viewed as virtuous (e.g., donating to St. Jude) and θ_A increases: then, signals from the H group provide less information and should be given less weight. Social sophistication suggests that a predictor should account for the relative magnitude of SDR_A across actions. Thus, as the perceived virtue or stigma of an action grows relatively more unanimous or extreme, sophisticated predictors should increase their discounting of H-group signals relative to IC-group signals.

Hypothesis 3 *(Relative magnitude).* Predictors with social sophistication will increase the relative weight given to incentive-compatible information as the perceived virtue of an action becomes more extreme.

In terms of our stylized model, this amounts to evaluating whether predictors are better than random at ordering SDR_A across actions.

4. Data description and manipulation checks

In this section, we provide summary statistics for each action in each part of the experiment. We then present our manipulation checks, demonstrating that (i) socially desirable responding is present and predictable and (ii) our predictors give positive weight to the signals they receive. Appendix Section B provides details on all of our estimation methods.

Our manipulation checks serve to establish that the bias from SDR is systematic and predictable. That is, differences between the IC and H groups are not exclusively attributable to noise or random choice errors. Absent this confirmation, no degree of social sophistication would allow a predictor to extract information from the statements of the H group that could improve their guesses about the choices of the IC group because no such information would exist. Thus, by confirming our manipulation checks, we confirm sufficient conditions that allow us to test for the presence of social sophistication.

Table 2 presents descriptive results for each action. The Sentiment Study and Choice Study are captured in Columns 1 and Columns 2–3, respectively. Initial and updated guesses from the Prediction Study are in Columns 4–6. Columns 7–8 compare predictors' average accuracy across information sources, where accuracy is measured by the absolute difference between a predictor's updated guess and the true value.[26]

4.1. Socially desirable responding

In order to conduct valid tests of social sophistication among predictors, we must first establish that SDR is present in the signals they receive. Recall that we defined SDR as the difference in selection rates between the H and IC groups ($SDR_A \equiv H_A - IC_A$). Thus, we must first ensure that the H group overstates (understates) claimed demand for socially desirable (undesirable) behaviors relative

[24] This hypothesis was not included in our pre-analysis plan. We include it here and provide results in Section 5 because they meaningfully add to our understanding of social sophistication among predictors. Our analysis faithfully replicates the analysis we used to evaluate every other hypothesis.

[25] This assumes that the predictor does not use the signal to reverse their opinion that the action is over-reported on average. If this were the case, then the predictor may conclude that donations to St. Jude are in fact under-reported on average. Results from the Sentiment Study support our assumption, showing that actions have predictable social desirability.

[26] With political behaviors that seem in conflict with each other, one concern is that distracted subjects may engage in both. We find this to be rare. In the 748 political decisions we observe (both hypothetical and incentivized), only 5 subjects selected both sides of contrary options: one subject donated to both Biden and Trump, two donated to both Trump and the DNC, and two donated to both the DNC and the RNC.

Table 2
Summary statistics for each action.

Action	Sentiment	Choice Rate		Initial	Updated Guesses		Updated ABS Error	
	(V_A)	IC Group	H Group	Guesses	IC Signal	H Signal	IC Signal	H Signal
St Jude Donation	9.15	61.5%	75.0%	60.8%	63.4%	68.8%	12.9	16.8
NPR Donation	6.14	26.4%	31.3%	26.5%	26.5%	26.8%	11.7	14.4
Steal	2.18	25.3%	19.8%	44.7%	30.9%	29.6%	13.5	15.1
Take Donation	4.07	12.1%	6.3%	24.1%	13.6%	11.1%	9.0	9.1
Trump Donation	3.80	11.0%	17.7%	30.1%	18.4%	21.9%	11.5	13.7
Biden Donation	3.74	3.3%	8.3%	24.2%	10.7%	11.0%	8.4	8.9
RNC Donation	4.72	7.7%	20.8%	33.1%	17.6%	24.5%	11.9	17.7
DNC Donation	4.53	12.1%	25.0%	33.6%	18.2%	26.3%	10.3	16.1

Notes: This table does not include data from subjects who were dropped from the analysis because of duplicate entries (see Section 2). Sentiment (V_A) is a within-subject average of three responses from 0 to 10 about the social desirability of the action.

Table 3
Socially desirable responding and perceived virtue.

	Socially Desirable Responding	
Mean Sentiment	2.390*	
	(1.12)	
Standardized Sentiment		3.112***
		(0.48)
Constant	-6.080	5.375***
	(5.814)	(0.00)
Observations	8	312
Clusters	N/A	39

Notes: "Mean Sentiment" aggregates 39 evaluations measured from 0 (Very Negative) to 10 (Very Positive). "Standardized Sentiment" normalizes sentiment ($V_{i,A}$) within each individual to have mean 0 and SD 1. Column 1 presents OLS results. Column 2 presents results of a linear regression with subject-level random effects and standard errors clustered at the subject level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

to the IC group. Additionally, the inflation of claimed demand for an action must not be random, but rather systematically tied to the action's perceived desirability, which we measured independently during the Sentiment Study.

Table 3 presents this analysis at two levels of specificity. Column 1 regresses SDR_A on V_A, the mean perceived virtue of the action from the Sentiment Study (see Equation (1)). Column 2 follows with an individual-level version of this test that regresses SDR_A on $\widehat{V}_{i,A}$, the within-subject normalized index of the action's perceived virtue (see Equation (2)).

Our measure of SDR is clearly predicted by the evaluations of social desirability from the Sentiment Study. Column 2 shows that the H group overstates their claimed demand for socially desirable behaviors by an additional 3.1 percentage points for every one standard deviation increase in perceived virtue.[27]

The fact that subjects' evaluations of sentiment are predictive of observed SDR demonstrates their "sentiment sophistication:" they have a fairly accurate understanding of the relative stigma or virtue surrounding an action.[28] Using the knowledge of which actions are more socially desirable—and therefore more likely to inspire dishonest responses from the H group—subjects could tailor their discounting of the H group's claims to control for SDR. Our tests of Hypotheses 2 and 3 evaluate whether predictors can complete this operation and translate knowledge of the social desirability of an action—obtained through sentiment sophistication—into knowledge of the resulting bias—a measure of social sophistication.

This sentiment sophistication is presented graphically in Fig. 3, which orders each of the eight actions along the horizontal axis according to their observed SDR. For each action, the associated sentiment evaluations are plotted on the vertical axis, revealing a clear positive association between the action's perceived virtue and the SDR in the Choice Study.

In Appendix Section A.1, we perform several other tests of our sentiment metrics to confirm their ability to capture social desirability. First, in Appendix Table A.2, we show that each of the three components of our sentiment index correlates with the other two components in predictable ways. Second, in Appendix Table A.3 we show that the variance in perceptions of social desirability shrinks as our questions about sentiment move from individual judgments toward estimates of social judgments. That is, our evaluators exhibit some disagreement in how they would *personally* judge certain actions but largely agree on how society more broadly would judge those actions. Finally, in Appendix Table A.4, we show that pairwise correlations between the sentiment evaluations of separate

[27] Appendix Table A.1 breaks down this association by each of the three components of our sentiment index. The relationships are similar across components, though others' sentiment and second-order perceptions of sentiment appear to be slightly stronger predictors of SDR than one's own sentiment.

[28] Note that our tests only compare *relative* sentiment and *relative* SDR across actions. Thus, the *absolute* measures of virtue or stigma do not have a clear interpretation.

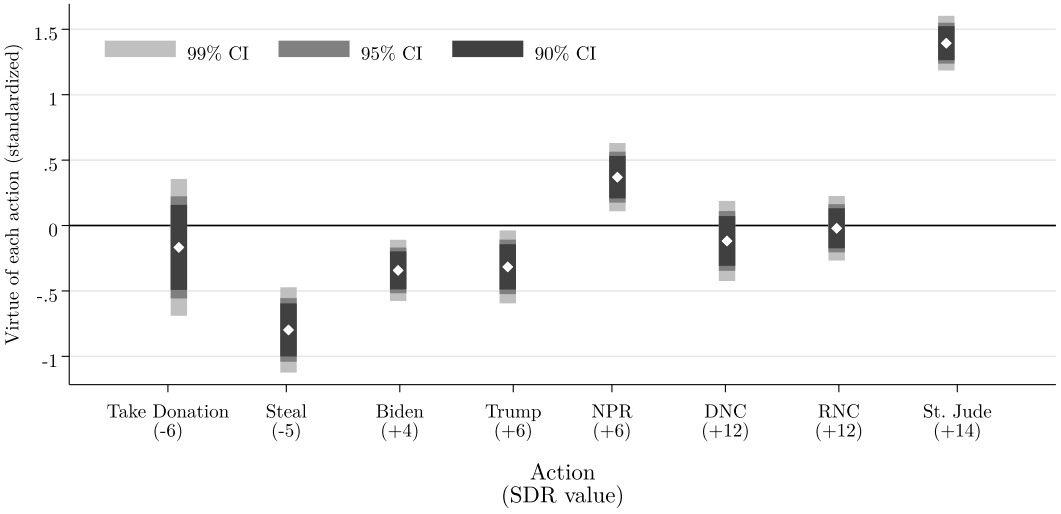

Fig. 3. Sentiment associated with each action. Actions ordered by SDR value.

Table 4
Improvements in accuracy depending on information source.

	Absolute Errors		Squared Errors	
	Updated Error	ΔError	Updated Error	ΔError
IC Info Source	-2.76***	-2.73***	-115.33***	-98.85
	(0.59)	(1.04)	(30.07)	(74.03)
Initial Error	0.24***		0.16***	
	(0.02)		(0.03)	
Constant	5.02***	-11.41***	113.80***	-576.66***
	(0.68)	(1.24)	(33.45)	(90.81)
Mean Initial Error:	21.58		792.99	
Standard Deviation:	(18.09)		(1219.52)	
Observations	2168	2168	2168	2168
Clusters	271	271	271	271

Notes: Linear regression with subject-level random effects. Standard errors clustered at the individual level. Fixed effects are included for each action. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

actions have the anticipated signs. For example, those who positively evaluate donations to the Republican National Committee tend to positively evaluate donations to Donald Trump as well. Thus, in addition to predicting biases from SDR, our sentiment evaluations pass key tests regarding their validity in capturing social sentiment.

4.2. Accuracy

All predictors were tasked with guessing the behavior of the IC group. Therefore, signals drawn from the choices of the IC group will necessarily be (weakly) more predictive than signals drawn from the claims of the H group. Thus, we can validate that predictors are responsive to signals by testing if higher-quality information (i.e., from the IC group) results in more accurate updated guesses.

Table 4 presents this manipulation check. Columns 1-2 measure accuracy based on the absolute error of a predictor's guess: $|IC_A - \text{GUESS}_{i,t,A}|$, where $t \in \{1,2\}$ denotes the initial and updated guess, respectively. Columns 3-4 repeat this analysis using the squared error of a predictor's guess: $\left(IC_A - \text{GUESS}_{i,t,A}\right)^2$. Baseline accuracy was balanced across predictors who were randomly assigned to receive signals from either the IC or H group.[29] Therefore, our outcome of interest is the extent to which predictors' updated guesses become more accurate depending on their information source.

Here, "IC Info Source" is an indicator variable equal to one if the predictor is assigned to receive signals from the IC group. Columns 1-3 show that receiving this higher-quality information causes a large and statistically significant improvement in the accuracy of predictors' guesses. That is to say, higher-quality signals lead to more accurate updated guesses.

[29] p-values for differences in baseline absolute- and squared-errors are $p = 0.97$ and $p = 0.81$, respectively.

It is important to note that the constant terms estimated in Columns 2 and 4 are negative and significant. Thus, on average, the error in a subject's guess decreases after receiving information, regardless of the information source. Even the lower-quality signals from the H group improve predictors' guesses relative to their initial accuracy.

Moreover, while we do not directly compare subjects' updated guesses to a Bayesian benchmark, subjects do tend to combine their initial guesses and signals in reasonable ways. In particular, 84% of updated guesses fall weakly between the initial guess and the signal. 82% of subjects provide updated guesses that exhibit this "betweenness" property for at least six of the eight actions, while 42% of them exhibit it for all eight actions.

Although updated guesses do improve in accuracy, predictors in both groups still fall short of a simple heuristic: completely following their signals. Both groups would improve their accuracy by simply making guesses that match their signals exactly.[30] On average, signals from the IC group have an absolute error of 8.6 percentage points, while the associated updated guesses have an absolute error of 11.1 (test of differences: $p < 0.001$). Signals from the H group have an absolute error of 12.2 percentage points, while the associated updated guesses have an absolute error of 13.9 (test of differences: $p < 0.001$).

5. Main results

Our manipulation checks confirmed that SDR is widespread and predictable and that predictors' guesses are sensitive to their signals. With these prerequisites established, we now proceed to test our hypotheses about social sophistication, exploring the extent to which predictors anticipate and react to SDR. Our analysis closely follows our pre-registration with few amendments. As we evaluate each hypothesis, we will begin with our pre-registered specification before presenting any alternative specifications. Appendix Section B details our empirical estimation and how we supplement our pre-registered analysis. For reference, Hypotheses 1 and 3 are included in the pre-registration, while Hypothesis 2 is not. However, we adopt the same statistical approach for all hypotheses. Sections 5.4 and 5.5 also adapt this statistical approach to perform exploratory analysis on predictor confidence and signal noise, respectively.

5.1. Hypothesis 1: anticipation of SDR

Hypothesis 1 states that social sophistication should cause predictors to give signals from the IC group relatively more weight than those from the H group, on average. This amounts to testing if predictors identify differences in information quality between the two sources and discount hypothetical claims relative to actual choices.

Evaluating a predictor's sensitivity to their idiosyncratic signals from either the IC or H group poses a particular obstacle: participants in the two groups faced different incentives in the Choice Study, and thus the distribution of signals differs across groups. Therefore, our random assignment of information source is confounded with the assignment of a different distribution of signals. To resolve this confound and isolate the random sampling variation in signals, we control for the different distributions from which signals are drawn.[31] We accomplish this by including either (i) controls for the mean of the signal distribution or (ii) fixed effects for the distribution. We then causally identify the *differential* impact of signals from the IC group because of our randomly-assigned information source (IC vs. H).

Table 5 presents our test of Hypothesis 1—whether predictors anticipate SDR and accordingly give greater weight to information from the IC group when updating their guesses. Column 1 follows our pre-registration exactly, estimating the updated guess while controlling for the initial guess with additional controls for the mean of the signal distribution. Column 2 examines within-predictor changes in guesses, which increases statistical power. Column 2 also employs a more conservative solution to address the differences in distributions by including fixed effects for each of the 16 combinations of actions and information sources. Columns 3 and 4 replicate the analysis of Column 2 but restrict our sample to newly recruited predictors and experienced predictors, respectively. This allows us to explore the role of personal experience in prompting skepticism toward claims from the H group. Finally, Column 5 restricts the sample to two actions with unambiguous moral valence—donations to St. Jude Children's Hospital and stealing from another subject. This restricted sample allows us to understand whether the weights assigned to specific signals in our full sample may be driven by ambiguous sentiment toward political actions.

Columns 1 and 2 of Table 5 reveal the skepticism with which predictors treat signals from the H group. Predictors respond to each mechanically-random one-percentage-point increase in a signal from the H group by updating their guesses by 0.55–0.59 percentage points ($p < 0.001$ for both)—about halfway to the signal. The interaction term "IC Info Source×Signal Value" shows that predictors give signals from the IC group significantly greater weight, confirming Hypothesis 1. A one-percentage-point increase in an IC-group signal results in a *greater* increase in a predictor's updated guess than an identical increase in an H-group signal. This difference is equal to 0.08–0.17 percentage points ($p < 0.01$ for both). Equivalently, 14–31% of the updating from IC-group signals is attributable to "extra updating" due to the added weight given to IC-group signals relative to H-group signals.

Concretely, an IC-group signal showing one additional person (out of the 10 sampled) choosing action A will cause a predictor to increase their guess by 6.7–7.2 percentage points. In contrast, had this signal arrived from the H group, predictors would only increase their guess by 5.5–5.9 percentage points.

In Appendix Section A.2, we plot each initial and updated guess individually and perform heterogeneity analysis to show that predictors discount signals from the H group at both the intensive and extensive margins. We find suggestive (but not significant)

[30] More specifically, if one's signal reveals that z out of 10 people took the action, then this strategy calls for a guess that the IC choice rate is $(10 \times z)$%.

[31] See Kahan (2015) and Thaler (2024) for discussions on why responses to information alone are insufficient to identify differential updating.

Table 5
Updated guesses in response to signals from different sources.

	Updated Guess	Δ Guess			
	Full Sample	Full Sample	New Predictors	Experienced Predictors	Only "Steal" & "St. Jude"
Signal Value	0.59***	0.55***	0.55***	0.54***	0.41***
	(0.03)	(0.05)	(0.07)	(0.06)	(0.09)
IC Info Source × Signal Value	0.08***	0.17***	0.11	0.22**	0.19
	(0.03)	(0.07)	(0.10)	(0.09)	(0.12)
Initial Guess	0.30***				
	(0.02)				
IC Info Source	-1.14				
	(1.07)				
Observations	2168	2168	760	1408	542
Clusters	271	271	95	176	271
Control for Mean Signal:	Yes	N/A	N/A	N/A	N/A
Fixed-Effects:	Action	Action × Source	Action × Source	Action × Source	Action × Source

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

evidence that predictors with signals from the H group are both more likely to entirely ignore their signals, and less likely to submit updated guesses that exactly match their signals.

Predictors who participated in the H group during the Choice Study may have experienced the temptation to distort their responses, making them more skeptical when receiving signals from the H group. Thus, we use Columns 3 and 4 of Table 5 to test if experience in the Choice Study is a source of skepticism toward H-group signals. Predictors who previously participated in the Choice Study give 0.22 percentage points extra weight to each percentage-point increase in IC-group signals relative to H-group signals. On the other hand, newly recruited predictors only give IC-group signals 0.11 percentage points extra weight for a corresponding increase in signals. The difference between these groups is not significant, though when each prior role—IC group or H group—is analyzed separately, the differential effect of participating in the H group during the Choice Study approaches marginal significance ($p = 0.125$). These results can be found in Appendix Section A.3.[32]

The restricted sample of Column 5 finds largely consistent results. When only considering donations to St. Jude Children's Hospital and stealing from another subject, the magnitude of added weight given to IC-group signals is unchanged. Since the sample size shrinks to one-fourth of the original size, the precision is necessarily weaker. The results nonetheless suggest that predictors' discounting of signals from the H group is present across all actions, regardless of whether these actions have obvious moral valence.

Our results from Table 5—along with our supplemental analysis in Appendix Section A.2—consistently find that predictors demonstrate a fundamental feature of social sophistication: they anticipate the potential for SDR and respond by discounting the claims of the H group. Full social sophistication, however, involves more complex procedures that we examine next.

5.2. Hypothesis 2: direction of SDR

Here, we test if the predictors' guesses account for the direction in which SDR will bias signals from the H group. We examine how accurately predictors recognize whether the H group tends to overstate or understate their desire to take a given action and whether predictors adapt their weighting of signals accordingly.

Predictors with social sophistication should discount H-group signals more when they are "SDR-congruent"—i.e., suggestive of more socially desirable behavior—because H-group signals are biased in this direction. Conversely, social sophistication leads predictors to give *more* weight to H-group signals when they are "SDR-dissonant"—i.e., suggestive of less socially desirable behavior—because when such a signal is drawn despite the H group's exaggerated claims, the actual choices of the IC group are likely even less desirable. For a concrete example and details on this logic, see the related discussion in Section 3.2. Additionally, note that SDR-congruent signals do not necessarily point toward "more virtuous" behavior in some absolute sense, only behavior further in the direction that is over-reported.

To evaluate this hypothesis, we must first designate which actions tend to be over-reported due to SDR. We do so empirically using SDR_A. Action A is over-reported when $SDR_A > 0$ (i.e., the H group overstates their demand for A) and under-reported when $SDR_A < 0$.[33] With knowledge of an action's social desirability, social sophistication will enable predictors to determine if the signal

[32] Table A.6 contains our pre-registered analysis of the role of experience on social sophistication. The results are qualitatively similar to those in Columns 3 and 4 of Table 5.

[33] Note that all of our actions have $SDR_A > 0$ except for stealing from another subject and taking money from the Make-A-Wish Foundation.

Table 6
Updating from SDR-congruent and -dissonant signals from different sources.

	SDR-Congruent			SDR-Dissonant		
	Updated Guess	Δ Guess		Updated Guess	Δ Guess	
Signal Value	0.54***	0.43***	0.43***	0.57***	0.54***	0.19
	(0.07)	(0.11)	(0.14)	(0.04)	(0.07)	(0.29)
IC Info Source X Signal Value	0.14*	0.34**	0.28	0.02	0.06	0.07
	(0.07)	(0.16)	(0.21)	(0.04)	(0.11)	(0.38)
IC Info Source	-3.65			1.64		
	(2.31)			(1.27)		
Initial Guess	0.31***			0.28***		
	(0.03)			(0.03)		
Observations	920	920	422	1248	1248	120
Clusters	271	271	264	271	271	113
Control for Mean Signal:	Yes	N/A	N/A	Yes	N/A	N/A
Fixed-Effects:	Action	Action×Source		Action	Action×Source	

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. "SDR-congruent" ("SDR-dissonant") is defined by whether the signal is in the direction of more (less) social desirability relative to the average initial guess for that action across all predictors. Columns 1-2 and Columns 4-5 use the full sample of SDR-congruent and SDR-dissonant signals, respectively. Columns 3 and 6 restrict the sample to only the actions "Steal" and "St. Jude" for SDR-congruent and SDR-dissonant signals, respectively. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

they receive is SDR-congruent or SDR-dissonant. As discussed in Section 3.2, an SDR-congruent signal is one that indicates greater demand for an action with $SDR_A > 0$ (or lesser demand for an action with $SDR_A < 0$) *than the predictor initially guessed*. An SDR-dissonant signal indicates lesser demand for an action with $SDR_A > 0$ (or greater demand for an action with $SDR_A < 0$) *than the predictor initially guessed.*

Note that universal thresholds for SDR-congruent (-dissonant) signals do not exist since they are defined with respect to a predictor's initial guess. Individual-level analysis of this kind is thus necessarily endogenous because it conditions on predictor characteristics. We overcome this endogeneity by employing a population-level definition of an SDR-congruent (-dissonant) signal based on whether it indicates more (less) socially-desirable behavior than the *average* initial guess across all predictors. This approach removes endogeneity but reduces precision as it leaves open the possibility that a given signal is categorized as SDR-congruent even though it falls below the initial guesses of some individuals. In Appendix Table A.7, we replicate this analysis under a more precise, individual-level categorization of signals that exhibits some of the endogeneity concerns.

Our test of Hypothesis 2 modifies the approach of Hypothesis 1 to test if the weight given to H-group signals depends on whether they are SDR-congruent or SDR-dissonant. To aid the interpretation of coefficients, we will replicate the analysis of Hypothesis 1 separately for predictors receiving SDR-congruent and SDR-dissonant signals.

Table 6 displays our limited support for Hypothesis 2. Columns 1–3 replicate the analysis of Columns 1, 2, and 5 from Table 5 but restrict their focus to SDR-congruent signals. In this direction, signals from the H group should be discounted relative to those from the IC group. We find a positive coefficient for "IC Info Source×Signal Value," revealing that H-group signals receive less weight than IC-group signals. For every one-percentage-point increase in signals, Column 1 shows that predictors' guesses increase by 0.14 percentage points less when the signals arrive from the H group ($p = 0.063$). Column 2 estimates this diminished weight to be 0.34 ($p = 0.032$). Thus, predictors do appear to recognize that SDR-congruent signals are less credible when they come from the H group instead of the IC group, though the statistical significance of this result depends on the specification. Column 3 shows that this behavior is similar in magnitude when only considering donations to St. Jude Children's Hospital and stealing from another subject, though the statistical significance is diminished in the smaller sample.

Social sophistication also requires recognizing that SDR-dissonant signals should be given relatively *more* weight when they come from the H group. Columns 4–6 show no evidence for this more complex dimension of social sophistication. Regardless of specification or sample, we find a small positive coefficient for "IC Info Source×Signal Value" instead of the predicted negative coefficient. While these estimates are not significant, the positive sign of these effects means that, if anything, predictors still discount signals from the H group even when they are SDR-dissonant. Thus, we conclude that predictors' guesses do not demonstrate a recognition that SDR-dissonant signals from the H group are even stronger indictments of behavior than corresponding signals from the IC group.[34]

[34] In Appendix Section A.4, we present individual guesses in Figs. A.3 and A.4 to visualize heterogeneous discounting with respect to the direction of SDR.

Table 7
Updated guesses in response to SDR magnitude by information source.

	Updated Guess	Δ Guess		
Signal Value	0.60***	0.57***	0.63***	0.27
	(0.05)	(0.12)	(0.06)	(0.27)
IC Info Source×Signal Value	0.06	0.38**	0.24**	0.61*
	(0.07)	(0.17)	(0.09)	(0.35)
Signal Value×$\lvert SDR\rvert$	-0.00	-0.00		0.01
	(0.00)	(0.01)		(0.02)
IC Info Source×Signal Value×$\lvert SDR\rvert$	0.00	-0.02		-0.04
	(0.01)	(0.02)		(0.03)
IC Info Source×$\lvert SDR\rvert$	-0.14	-0.46		-0.14
	(0.16)	(0.47)		(0.73)
$\lvert SDR\rvert$	0.17	-0.40		-1.91
	(0.13)	(0.90)		(0.55)
Initial Guess	0.32***			
	(0.02)			
IC Info Source	0.19			
	(1.58)			
Signal Value×$\lvert\widehat{V}_A\rvert$			-0.10**	
			(0.05)	
IC Info Source×Signal Value×$\lvert\widehat{V}_A\rvert$			-0.04	
			(0.07)	
IC Info Source×$\lvert\widehat{V}_A\rvert$			-4.66	
			(5.20)	
$\lvert\widehat{V}_A\rvert$			35.81	
			(57.97)	
Observations	2168	2168	2168	542
Clusters	271	271	271	271
Control for Mean Signal:	Yes	N/A	N/A	N/A
Fixed-Effects:	Action	Action × Source	Action × Source	Action × Source

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. Columns 1–3 feature the full sample. Column 4 restricts the sample to the actions: "Steal" and "St. Jude." * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.3. Hypothesis 3: relative magnitude of SDR

We now test if predictors appreciate which claims from the H group are more susceptible to SDR and, therefore, more worthy of discounting. Table 3 and Fig. 3 show how Sentiment Study responses can predict which actions generate the strongest image concerns. Here, we examine if predictors apply such knowledge when interpreting claims from the H group.

Our test of Hypothesis 3 adapts the approach of Hypothesis 1 to include interaction terms for the level of SDR_A. As SDR_A grows in magnitude, H-group claims are increasingly distorted by SDR and social sophistication prescribes greater discounting for such claims. Since the discounting of H signals should increase with the magnitude of SDR_A regardless of its sign, we use its absolute value, $|SDR_A| = |H_A - IC_A|$, as our interaction term.

To demonstrate robustness to an alternative measure of SDR, and to directly connect social sophistication with sentiment sophistication, we repeat the analysis above using responses from the Sentiment Study as the interaction term. Table 3 and Fig. 3 confirm that SDR tends to increase in magnitude as the sentiment group's evaluations of an action become more extreme; thus, predictors should increasingly discount signals from the H group for such actions. Our interaction term in this case is the absolute value of a normalized measure of sentiment at the action-level: $|\widehat{V}_A| = \left|\frac{V_A - \overline{V}}{\sigma_V}\right|$, where $\overline{V}$ and σ_V are the mean and standard deviation of V_A (Equation (1)) across all eight actions.

Table 7 presents our test of Hypothesis 3. We find no evidence that predictors increase their relative discounting of signals from the H group as either SDR or perceived virtue become more pronounced. In Columns 1, 2, and 4 the additional discounting of the H group is captured by the coefficient for "IC Info Source×Signal Value×$|SDR|$." We find no evidence of increased discounting of H-group claims for actions with greater SDR. In fact, we find point estimates in the wrong direction. In Column 3, the additional discounting of the H group is captured by the coefficient for "IC Info Source×Signal Value×$|\widehat{V}_A|$." As in Columns 1 and 2, predictors fail to increase their discounting of claims from the H group as $|\widehat{V}_A|$ grows, with point estimates again in the wrong direction. Thus, Table 7 rejects the notion that predictors tailor their inferences to the relative magnitude of bias from SDR.

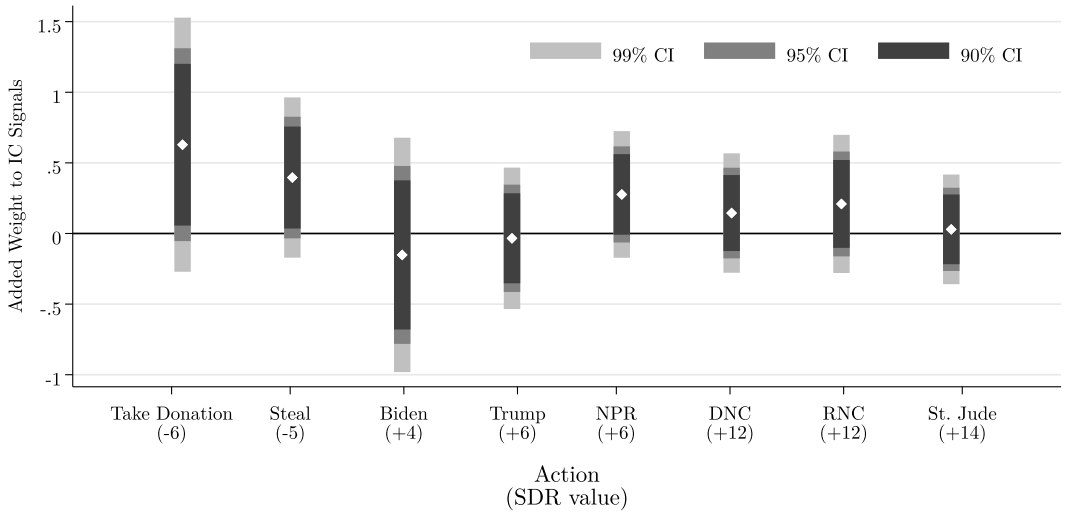

Fig. 4. Weight given to signals from the IC group. Actions ordered by SDR value.

Fig. 4 visualizes this lack of social sophistication.[35] As in Fig. 3, actions are ordered by SDR_A, with extreme negative values on the left and extreme positive values on the right. With social sophistication, the relative weight given to IC signals (and hence the relative discounting of H signals) should grow at the extremes. We find no such pattern. In fact, for the action with the most extreme SDR—donations to St. Jude Children's Hospital—predictors do not discount H signals at all. Thus, in contrast to Fig. 3—which demonstrated clear sentiment sophistication—Fig. 4 finds no evidence of social sophistication.

Recall from our discussion in Section 3 that Hypothesis 3 provides a rather weak test of social sophistication—it amounts to testing whether predictors' guesses account for the relative SDR across actions in a way that is better than random. As is evident from Fig. 4, predictors fail this test. This failure is striking because the test is a natural extension of the tests from Table 3 and Fig. 3. In those, the sentiment group demonstrates an understanding of which actions tend to be socially desirable. Thus, it appears that predictors fail to translate the sentiment sophistication that is clearly present in the population into the discounting behaviors prescribed by social sophistication.

5.4. Confidence in predictions

Immediately after making a guess, we asked predictors to state their confidence in that guess on a scale from 0 to 10. Although these elicitations were not incentivized, they provide further insight into the perceived differences between the two information sources. Table 8 examines the association between confidence and the absolute error of a guess. We specifically focus on how higher-quality information from the IC group influences this relationship. Since IC-group signals are weakly more informative, socially-sophisticated predictors who appreciate this fact should display greater increases in confidence when they receive information from the IC group. With our random assignment of the information source, we can causally identify the relationship between higher-quality information and confidence in predictions.

Our analysis uses absolute errors to measure accuracy, meaning that positive numbers indicate diminished accuracy. Initial confidence and updated confidence are both normalized across all individuals and actions to have a mean of 0 and a standard deviation of 1.

Column 1 shows that predictors have a false sense of confidence. Similar to the classic result from Kruger and Dunning (1999), there is a positive and significant relationship between the error in predictors' initial guesses and their initial confidence ($p < 0.01$). Column 2 shows that this false-confidence effect persists for the updated guesses of predictors who receive information from the H group ($p < 0.05$). However, the positive association between errors and confidence is diminished and statistically insignificant for predictors who receive higher-quality information from the IC group ($p > 0.40$).[36] Interestingly, receiving information from the IC group has a near-zero and insignificant level effect on confidence, despite the greater accuracy. This shows a clear failure to notice the material difference in the informativeness of the two sources.

Taken together, these results suggest not only that predictors fail to account for biased claims from the H group, but they also fail to notice key differences between these information sources. These failures may drive second-order consequences such as the persistence of unfounded confidence in erroneous predictions. Moreover, this casts particular doubt on the assertion that predictors discount claims of the H group because they believe this information is noisier.

[35] Fig. 4 presents coefficients and confidence intervals for "IC Info Source×Signal Value" separately for each action. The specification is drawn from Column 2 of Table 5 and replicated for each individual action. We include an indicator variable for "IC Info Source," since we cannot include fixed effects for each combination of action and information source. All regressions cluster standard errors at the subject level.

[36] Serra-Garcia and Gneezy (2021) find similar overconfidence in one's ability to detect lies by others.

Table 8
Confidence in predictions.

	Initial Confidence	Updated Confidence
Initial Error (Absolute Value)	0.004***	-0.006***
	(0.001)	(0.001)
Updated Error (Absolute Value)		0.004**
		(0.002)
Updated Error×IC Info Source		-0.002
		(0.003)
Initial Confidence		0.457***
		(0.023)
IC Info Source		-0.015
		(0.077)
Constant	-0.298***	0.119*
	(0.070)	(0.062)
Observations	2168	
Clusters	271	
Fixed-Effects:	Action	Action

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. Confidence is normalized to mean 0 and standard deviation of 1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.5. Noise and polarization

As noted above, our results on updating could be driven in part by predictors believing that H-group signals are less reliable due to inattentive hypothetical responses or uncertainty about (or polarization of) the social desirability of the action in question. Our manipulation checks in Section 4 contradict the notion that the claims of the H group are imprecise but unbiased, showing instead that they have predictable biases and do carry information. Nonetheless, predictors could hold exaggerated perceptions of the noise in H-group signals or believe them to arise from a source other than SDR. This may lead them to discount H-group signals even absent any concerns about SDR. Thus, some fraction of what we attribute to social sophistication about SDR could instead be driven by beliefs about noise. Our estimates would then represent an upper bound of possible sophistication. This alternative explanation not only contradicts our manipulation checks, as mentioned, but also contradicts our results about predictor confidence. If predictors believed H-group signals to be substantially noisier than IC-group signals because of pure, "white noise," they should have less confidence in their updated guesses after receiving H signals. As we showed in Section 5.4, predictors do not demonstrate these patterns of confidence in their updated guesses.

Predictors could also believe H-group signals contain noise resulting from uncertainty about or polarization towards the social desirability of an action. To further examine the potential effects of this form of noise on the discounting behavior of predictors, we construct a proxy for this form of noise that uses the variance in sentiment scores assigned to each action by the sentiment group. Notice that this measure also captures the "polarization" of an action—some actions (e.g., supporting a particular political candidate) may be viewed as virtuous by some yet taboo by others. Such polarization is another potential reason for predictors to discount H signals if it renders predictors uncertain about the extent of SDR. Fig. 5 orders our actions based on this variance measure and plots the added weight given to IC signals for each action. Aside from the extreme outlier of "Take Donation"—which has nearly three times the variance of the second highest variance action—there is no clear link between discounting behavior and the variance in sentiment.[37]

This variance measure also has an interesting connection to the confidence that predictors show in their guesses. If predictors were discounting H-group signals not because of perceived biases but because of uncertainty or noise, then we would expect them to have less confidence in their guesses about actions that have higher variance in sentiment. We find the opposite. Column 1 of Table A.8 in Appendix Section A.5 shows that predictors grow more confident in their initial guesses as the variance grows ($p < 0.001$). Column 2 shows that this does not persist for the updated guesses. But, importantly, the interaction term shows that predictors exhibit no added confidence from receiving the higher-quality IC-group signals when the sentiment variance is high, casting doubt on the idea that they may be concerned with unreliable or noisy signals. Column 3 shows that initial confidence does not derive from initial accuracy, as the accuracy of initial guesses tends to fall as the variance in sentiment grows ($p = 0.058$).

While our proxy variable for polarization and noise may not perfectly capture either phenomenon, it does provide strong evidence that these were not the primary concerns driving predictors' discounting behaviors. Moreover, the final columns of Tables 5–7 restrict our analysis to the two actions with the most obvious and agreed-upon sentiment—donating to St. Jude and stealing—and find similar patterns of updating relative to the full sample. In combination with the patterns of confidence our predictors showed, we believe

[37] We believe the variance in sentiment scores for "Take Donation" may result from confusion, because it is the only action that received both the highest (fully virtuous) and lowest (fully stigmatized) scores from different subjects in the Sentiment Study.

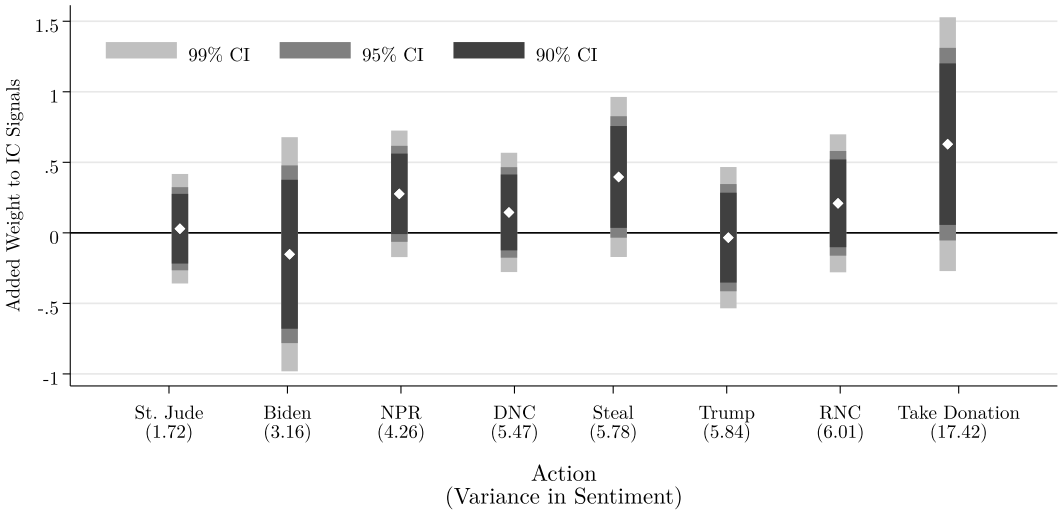

Fig. 5. Weight given to signals from the IC group. Actions ordered by variance in sentiment index.

there is sufficient evidence to rule out the possibility that predictors were actually sophisticated but worried about noise in the signals they received.

6. Discussion and conclusion

In our experiment, we designed an environment to cleanly identify SDR across several different actions. We then asked subjects to predict choice behavior for the actions. We presented subjects with random subsamples of data from either incentivized choices or unincentivized polls to assist them in their predictions. This novel subsampling approach offers a cleaner causal identification of responses to information than the traditional paradigm of information-revelation or belief-correction experiments. The traditional approach reveals identical information to all subjects, meaning that the direction of updating is endogenous to prior beliefs. We believe our approach can alleviate these endogeneity concerns and could even be adapted to other settings testing perceptions about the quality of information.

When our subjects were presented with data from unincentivized polls, they showed limited "social sophistication" in controlling for the SDR manifest in those poll data. Our subjects correctly put less weight on what others claimed they would do relative to what others actually did. However, they faltered in calibrating their discounting to the SDR of each action. Despite other subjects from the same population showing an ability to identify the social desirability of actions, predictors failed to translate this knowledge into sophisticated discounting. While our subjects correctly discounted SDR-congruent signals, they failed to give sufficient weight to SDR-dissonant signals: these signals are especially informative since few people will lie to make themselves look less socially desirable. Further, when considering actions with more extreme social desirability—which inspired more dishonest hypothetical claims—subjects did not increase their discounting.

Our setting was designed to maximize the control over outside variables in order to cleanly identify biases from SDR. In such an abstract environment where subjects are carefully observed, we might expect that biased reporting due to SDR would be relatively salient. In light of this, the limited evidence we find for social sophistication among predictors is even more striking. We should be skeptical of how well peoples' inferences will control for more subtle forms of SDR in natural settings if they do not account for the blatant SDR in our contrived environment. However, further research is needed to determine the impact of contextual factors on social sophistication.

Other notions of "sophistication" in behavioral economics typically require the recognition and anticipation of one's own biases. Such sophistication is rare (Heidhues and Kőszegi, 2010; Ericson, 2011; Augenblick and Rabin, 2019). Although social sophistication in our setting does not require any self-reflection—it only requires participants to recognize that *others* may succumb to social desirability bias—we still find limited evidence of sophistication.[38]

Future research can delve deeper into the paradigm of social sophistication that we introduce. An important next step is to disentangle how much of the failure of updating is attributable to a failure of social sophistication and how much is attributable to other forms of signal reliability (e.g. noise, polarization, etc.). We address this by constructing a proxy for signal reliability, but our approach relies on the assumption that predictors treat noise from different sources similarly. A future design could bring additional clarity by replicating our approach—predicting IC-group behavior based on IC- and H-group signals—and comparing the observed updating to an inverted approach—predicting H-group behavior based on IC- and H-group signals.

[38] A literature on "bias blind spots" finds that people possess a greater ability to recognize others' biases than their own (Pronin et al., 2002; West et al., 2012). Fedyk (2018) demonstrates this asymmetry in the domain of intertemporal choice.

A failure to correct for biases from SDR has significant economic costs. Election results, public-health issues, job-market forecasts, and social-policy preferences are all frequently predicted using unincentivized poll data that are susceptible to SDR.[39] Our study demonstrates systematic failures in the interpretation of such poll data. Although our poll data should not be interpreted at face value, we find that people do not exhibit the social sophistication necessary to de-bias the data. One possible solution could be to simply collect evaluations of sentiment surrounding different poll responses. This could improve forecasts by leveraging the association between sentiment data and bias. In this way, biased poll data may be adjusted before it can carry over into biased inferences. Another solution could involve "debiasing" observers of poll data by explicitly reminding them about the potential difference between hypothetical and incentivized choices. We took care to avoid drawing subjects' attention to this difference in order to analyze how well they accounted for it on their own. Future research could examine the extent to which social sophistication is improved when subjects are specifically cued to consider how image concerns may drive a wedge between hypothetical and incentivized decisions.

Declaration of competing interest

The authors declare that they have no relevant or material financial interest that relate to the research described in this paper, nor do they have any relationship with interested parties who may have material financial interest that relates to the research described in this paper.

Appendix A. Supplemental analyses

A.1. Further analyses of sentiment measures

Table 3 captures the relationship between SDR and our sentiment index, which is constructed by taking the mean of the three measures of sentiment listed below. In this section, we replicate the analysis of Table 3 after breaking down our sentiment index into these component parts. Below, Table A.1 explores the association between SDR and each of the following sentiment measures:

1. How would you feel about taking this action yourself?
2. How would you feel about other people who take this action?
3. How do you think most other people would feel about people who take this action?

For each action A, let $Q_{i,j,A}$ denote subject i's response to question $j \in \{1,2,3\}$ above. For each of these three measures, we regress SDR_A on the sentiment rating averaged over individuals, $\bar{Q}_{j,A} \equiv \frac{\sum_{i=1}^{N_S} Q_{i,j,A}}{N_S}$. The results of these regressions are reported in Columns 1, 3, and 5 of Table A.1. We also regress SDR_A on these same sentiment measures after standardizing them within an individual; that is, we regress SDR_A on $\hat{Q}_{i,j,A} \equiv \frac{Q_{i,j,A} - \bar{Q}_{i,j}}{\sigma_{i,j}}$, where $\bar{Q}_{i,j}$ and $\sigma_{i,j}$ are subject i's mean and standard deviation of $Q_{i,j,A}$ for measure j across all eight actions. The results of these regressions are reported in Columns 2, 4, and 5 of Table A.1. Note that the column headers (e.g., "Measure 1") in Table A.1 indicate which of the three questions above are used to form the regressor.

From these results, we can see consistent relationships between different measures of stigma and the observed socially desirable responding in the Choice Study. While these relationships are all positive and most are significant, there appears to be a stronger association with others' sentiment toward others (Columns 5–6) rather than sentiment toward one's self (Columns 1–2) or sentiment toward others (Columns 3–4). This would suggest that people may be more worried about the virtue or stigma they think others will attach to an action rather than the virtue or stigma they attach to the item themselves, though this would need more targeted research to confirm.

A.2. Heterogeneous responses to signals

The analysis in Table 5 is limited to aggregate updating and could obscure important heterogeneity in updating behavior. Figs. A.1 and A.2 add detail to explore this heterogeneity. Each figure shows all predictors' guesses relative to the signal they received. The x-axis (y-axis) measures the difference between a predictor's initial (updated) guess and her signal. Since a steeper slope indicates less weight given to the signal, our test of Hypothesis 1 from Table 5 amounts to testing whether the slope is flatter in Fig. A.1.[40] These figures demonstrate more subtle responses to signals as well. A predictor who entirely ignores the signal will land on the 45-degree line, while a predictor who fully updates her prediction to match her signal will land on the x-axis. Table A.5 tests whether these behaviors—in addition to partial updating—differ across information sources.

Table A.5 shows that, when a signal comes from the IC group, predictors are 2.7 percentage points less likely to completely ignore it ($p = 0.175$) and 3.2 percentage points more likely to match it exactly ($p = 0.153$). Column 3 shows that predictors who neither

[39] Polls are used to determine candidate viability and access to debate stages (Fox News, 2016), they influence voter turnout (Großer and Schram, 2010; Agranov et al., 2018; Bursztyn et al., 2024) and reported preferences (Cantú and Márquez, 2021), affect campaign contributions (Adkins and Dowdle, 2002), and may help entrench illiberal regimes (Carlson, 2018). Boukouras et al. (2023) find that, even in abstract environments, biased polls inhibit objective evaluation of candidates and shift electoral outcomes. The influence of polls is so significant that a market has arisen for "fake polls" that manipulate asset prices (Yeargain, 2020). In this way, polling biases have economic costs even absent any biases in how individuals consume and interpret them.

[40] This holds for the region above the x-axis. Below the x-axis would indicate an *overreaction* to the signal.

Table A.1
Socially desirable responding and perceived virtue.

	Sentiment toward Self		Sentiment toward Others		Others' Sentiment toward Others	
	Socially Desirable Responding					
Mean Sentiment	2.030		2.434**		2.504*	
	(1.28)		(0.98)		(1.10)	
Standardized Sentiment		2.156***		3.237***		3.489***
		(0.48)		(0.47)		(0.43)
Constant	-3.448	5.375***	-7.058	5.375***	-6.944	5.375***
	(6.06)	(0.00)	(5.39)	(0.00)	(5.83)	(0.00)
Observations	8	312	8	312	8	312
Clusters	N/A	39	N/A	39	N/A	39

Notes: "Mean Sentiment" is aggregated across 39 individual evaluations measured from 0 (Very Negative) to 10 (Very Positive). "Standardized Sentiment" normalizes sentiment ($V_{i,j,A}$) within each individual to have mean 0 and SD 1. For each of our three sentiment measures, the first column presents OLS results. The second column presents results of a random-effects linear regression with subject-level random effects and standard errors clustered at the subject level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2
Correlation between three sentiment measures.

	Steal	St. Jude	NPR	Trump	Biden	RNC	DNC	Take Donation
Panel A:	Sentiment toward Others							
Sentiment toward Self	0.725***	0.534**	0.404***	0.888***	0.780***	0.793***	0.674***	0.908***
	(0.10)	(0.25)	(0.12)	(0.06)	(0.11)	(0.09)	(0.09)	(0.05)
Constant	0.119	4.700*	4.628***	1.036***	2.040***	1.884***	2.648***	0.473*
	(0.17)	(2.36)	(0.80)	(0.35)	(0.48)	(0.59)	(0.55)	(0.27)
Observations	39	39	39	39	39	39	39	39
Panel B:	Others' Sentiment toward Others							
Sentiment toward Self	0.451***	0.462***	0.280**	0.076	0.046	0.129	0.203	0.834***
	(0.08)	(0.15)	(0.13)	(0.11)	(0.10)	(0.10)	(0.12)	(0.08)
Constant	1.033***	5.032***	4.757***	3.478***	4.183***	4.284***	4.107***	0.767*
	(0.28)	(1.45)	(0.83)	(0.56)	(0.41)	(0.54)	(0.55)	(0.40)
Observations	39	39	39	39	39	39	39	39
Panel C:	Others' Sentiment toward Others							
Sentiment toward Others	0.464***	0.488***	0.762***	0.106	0.120	0.207*	0.347***	0.951***
	(0.14)	(0.10)	(0.10)	(0.11)	(0.12)	(0.11)	(0.13)	(0.03)
Constant	1.262***	4.523***	1.077*	3.302***	3.806***	3.749***	3.078***	0.201
	(0.33)	(0.99)	(0.62)	(0.67)	(0.58)	(0.67)	(0.74)	(0.20)
Observations	39	39	39	39	39	39	39	39

Notes: Each panel presents the regression of one of our sentiment measures on another. All results are derived from OLS regressions with heteroskedasticity-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3
Variance of each sentiment measure across actions.

	Steal	St. Jude	NPR	Trump	Biden	RNC	DNC	Take Donation
Sentiment toward Self	9.993	3.115	8.973	11.888	6.406	12.239	10.868	19.632
Sentiment toward Others	7.178	2.147	4.904	11.028	7.204	11.099	8.779	17.726
Others' Sentiment toward Others	4.660	1.904	4.722	4.406	2.534	3.572	4.397	17.305
Sentiment Index	5.777	1.718	4.262	5.840	3.161	6.009	5.466	17.416
Observations	39	39	39	39	39	39	39	39

Notes: Each panel presents the regression of one of our sentiment measures on another. All results are derived from OLS regressions with heteroskedasticity-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

completely ignore their signal nor match their signal exactly continue to discount signals from the H group by 17 percentage points relative to the IC group ($p = 0.023$).

Table A.4
Pairwise correlation between sentiment index across actions.

	Steal	St. Jude	NPR	Trump	Biden	RNC	DNC	Take Donation
Steal	1.000							
St.Jude	-0.294 (p = 0.069)	1.000						
NPR	-0.044 (p = 0.792)	0.411 (p = 0.009)	1.000					
Trump	0.066 (p = 0.690)	0.156 (p = 0.343)	0.009 (p = 0.956)	1.000				
Biden	0.247 (p = 0.130)	-0.022 (p = 0.892)	0.055 (p = 0.738)	-0.182 (p = 0.267)	1.000			
RNC	-0.106 (p = 0.521)	0.264 (p = 0.104)	0.130 (p = 0.429)	0.796 (p = 0.000)	-0.047 (p = 0.778)	1.000		
DNC	0.154 (p = 0.351)	0.045 (p = 0.786)	0.220 (p = 0.179)	-0.352 (p = 0.028)	0.598 (p = 0.000)	-0.120 (p = 0.468)	1.000	
Take Donation	0.024 (p = 0.886)	-0.174 (p = 0.289)	0.056 (p = 0.734)	0.114 (p = 0.489)	-0.009 (p = 0.955)	0.231 (p = 0.158)	-0.155 (p = 0.346)	1.000

Notes: P-values presented in parentheses below correlational estimates.

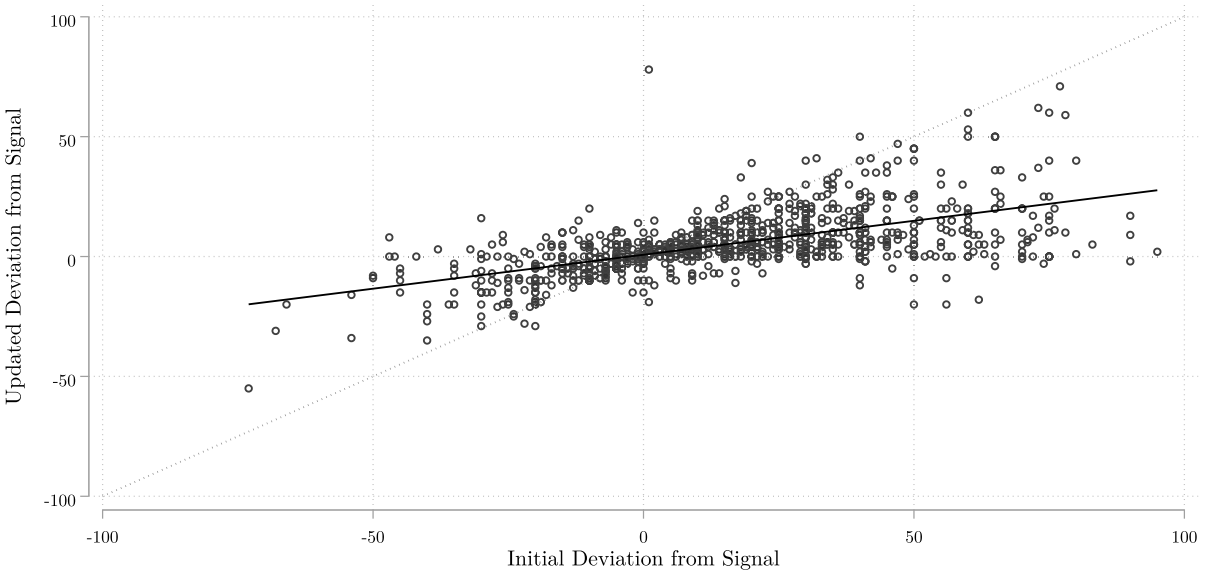

Fig. A.1. Predictors receiving signals from the IC group.

Table A.5
Updated guesses in response to signals from different sources.

	Pr[Ignore Signal]	Pr[Match Signal]	Δ Guess (partial updating)
IC Info Source	-0.03 (0.02)	0.03 (0.02)	
Signal Value			0.61*** (0.05)
IC Info Source×Signal Value			0.17** (0.07)
Observations	2168	2168	1663
Clusters	271	271	268
Fixed Effects:	None	None	Action×Source

Notes: Columns 1-3: Random-effects linear regression with subject-level random effects and standard errors clustered at the individual level. Column 3 restricts the sample to predictors who neither ignore their signal nor match their signal exactly. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

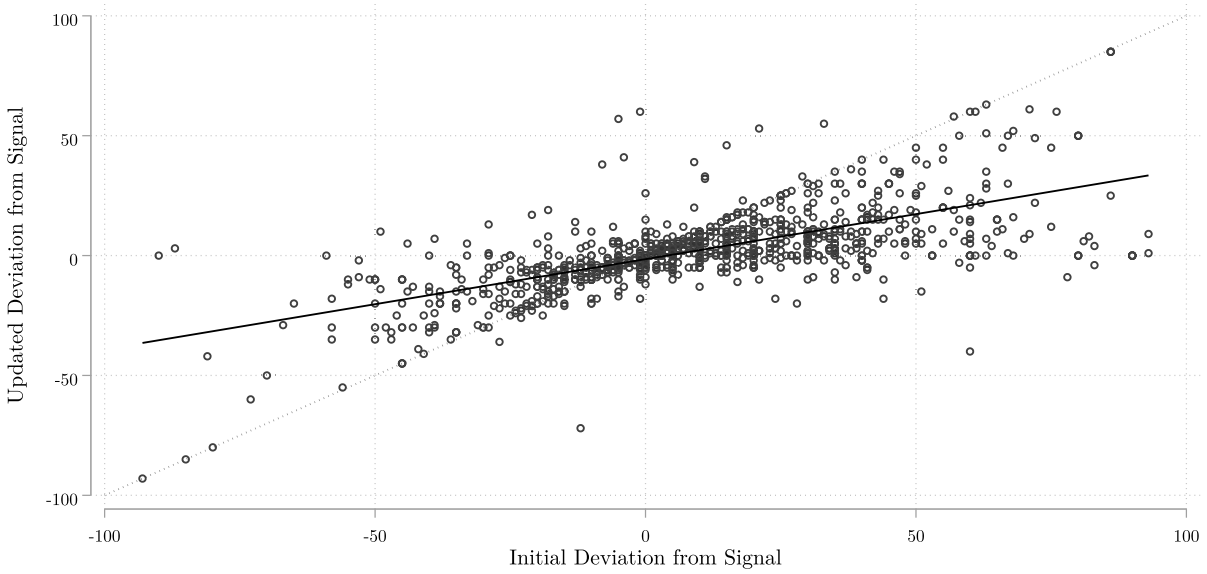

Fig. A.2. Predictors receiving signals from the H group.

A.3. Experience with SDR

To examine the mechanisms driving social sophistication, we explore whether predictors who previously participated in the Choice Study are better at accounting for SDR in poll data than newly recruited predictors. A predictor with experience in the Choice Study may have felt the impulse to misrepresent their own preferences. This experience may then be transformed into a higher degree of skepticism about signals from the H group. As a natural extension, we also test if this experience makes predictors more accurate in their guesses.

We specifically examine if the discounting of H signals relative to IC signals differs between three types of predictors: (i) those who participated in the IC group in the Choice Study, (ii) those who participated in the H group in the Choice Study, and (iii) newly recruited predictors who did not participate in the Choice Study. To test this, we adapt the approach of Hypothesis 1 to include interaction terms for each of the three groups.

Our results find no significant heterogeneity in the discounting of H signals relative to IC signals. Indeed, a fundamental level of social sophistication seems to be present in all predictors, including those who are newly recruited. However, there is some suggestive evidence that participants from the H group may give greater weight to IC signals. We find a positive point estimate of 0.12 ($p = 0.125$) for the coefficient on "IC Info Source×Signal Value×H Group Member". These predictors, having participated in the H group, may be more aware of the impulse to lie in the hypothetical Choice Study since they themselves faced this temptation. As a result, they may increase the relative weight they put on choices from the IC group, but this is speculative. This heterogeneity appears to grow in magnitude and significance when we limit the sample to the actions, "Steal" and "St. Jude." In this sample, predictors who participated in the H Group more than double the added weight given to IC-group signals relative to their newly-recruited peers ($p = 0.074$). These results are consistent, but merely suggestive and should be investigated further.

We also find no significant differences in the accuracy of predictors' guesses based on their experiences. The average absolute errors in first guesses are 21.54, 21.66, and 21.54 for predictors from the IC group, H group, and new recruits, respectively (joint test of equality $p = 0.99$). The corresponding average absolute errors in second guesses are 12.22, 12.61, and 12.58 (joint test of equality $p = 0.85$).

A.4. Direction of SDR

In Table A.7, we replicate the analysis of Section 5.2 using a more individualized measure of SDR-congruent and SDR-dissonant. In this case, we define each term based on whether the signal received indicates more or less socially desirable behavior *than the predictor initially guessed.* This introduces some endogeneity by relying on characteristics of the predictor to categorize the signals, but it does increase the precision of our estimation.

We again find that predictors discount SDR-congruent signals from the H group. However, the failure to give appropriate weight to SDR-dissonant signals from the H group also remains. Thus, under this specification, behavior is no more consistent with social sophistication than in Table 6.

Figs. A.3 and A.4 replicate the visualizations from Figs. A.1 and A.2 after replacing predictions about the number of subjects taking an action with the number of subjects engaging in the socially-desirable behavior. For example, this transformation replaces predictions about the number of subjects who steal from another subject with the number of subjects who refuse to steal from another subject.

Table A.6
Updated guesses by information source and prior experience.

	Updated Guess	Δ Guess	
Signal Value	0.59***	0.56***	0.45***
	(0.05)	(0.06)	(0.12)
Signal Value×IC Group Member	0.05	0.05	0.06
	(0.05)	(0.06)	(0.11)
Signal Value×H Group Member	-0.04	-0.07	-0.13
	(0.05)	(0.06)	(0.09)
IC Info Source×Signal Value	0.06	0.13*	0.11
	(0.05)	(0.08)	(0.15)
IC Info Source×Signal Value×IC Group Member	-0.02	0.01	-0.11
	(0.07)	(0.08)	(0.18)
IC Info Source×Signal Value×H Group Member	0.07	0.12	0.27*
	(0.07)	(0.08)	(0.15)
Initial Guess	0.30***		
	(0.02)		
IC Info Source	0.85		
	(1.98)		
IC Info Source×IC Group Member	-0.94	-0.33	5.38
	(2.54)	(4.16)	(9.93)
IC Info Source×H Group Member	-4.97*	-4.58	-13.29
	(2.69)	(4.03)	(9.32)
Observations	2168	2168	542
Clusters	271	271	271
Control for Mean Signal:	Yes	N/A	N/A
Control for IC/H/New Group:	Yes	Yes	Yes
Fixed-Effects:	Action	Action×Source	

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. Columns 1–2 feature the full sample. Column 3 restricts the sample to the actions: "Steal" and "St. Jude." * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7
Updating from SDR-congruent and -dissonant signals from different sources.

	SDR-congruent			SDR-dissonant		
	Updated Guess	Δ Guess		Updated Guess	Δ Guess	
Signal Value	0.61***	0.26***	0.13	0.71***	0.33***	0.23**
	(0.05)	(0.07)	(0.11)	(0.03)	(0.06)	(0.11)
IC Info Source×Signal Value	0.11***	0.12	0.23	-0.01	0.08	-0.00
	(0.04)	(0.09)	(0.14)	(0.03)	(0.09)	(0.15)
IC Info Source	-2.06			0.67		
	(1.27)			(1.28)		
Initial Guess	0.29***			0.23***		
	(0.04)			(0.03)		
Observations	925		352	1243		190
Clusters	267		242	270		161
Control for Mean Signal:	Yes	N/A	N/A	Yes	N/A	N/A
Fixed-Effects:	Action	Action×Source		Action	Action×Source	

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. "SDR-congruent" ("SDR-dissonant") is defined by whether the signal is in the direction of more (less) social desirability relative to the initial guess. Columns 1-2 and Columns 4-5 use the full sample of SDR-congruent and SDR-dissonant signals, respectively. Columns 3 and 6 restrict the sample to only the actions "Steal" and "St. Jude" for SDR-congruent and SDR-dissonant signals, respectively. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figs. A.3 and A.4 corroborate Table 6 by demonstrating a relatively similar response to information from the IC and H groups when that information suggests less socially-desirability (on the right side of the figures) and a significant discounting of information from the H group when that information suggests greater social-desirability (on the left side of the figures). This can be seen by the flatter slope on the left side of Fig. A.3 than on the left side of Fig. A.4.

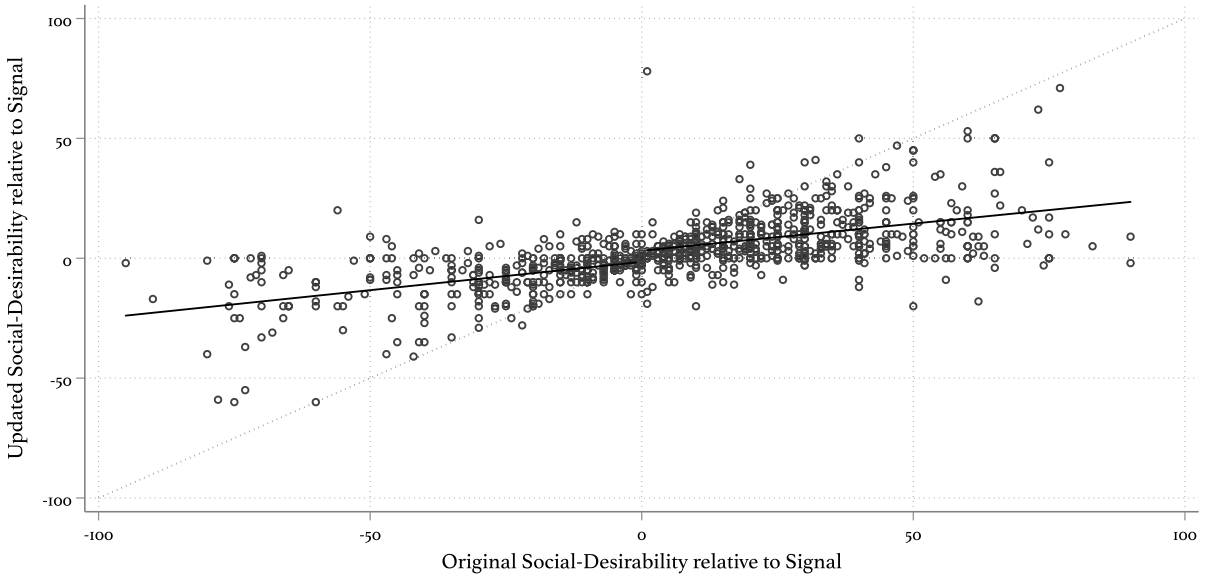

Fig. A.3. Predictors receiving signals from the IC group.

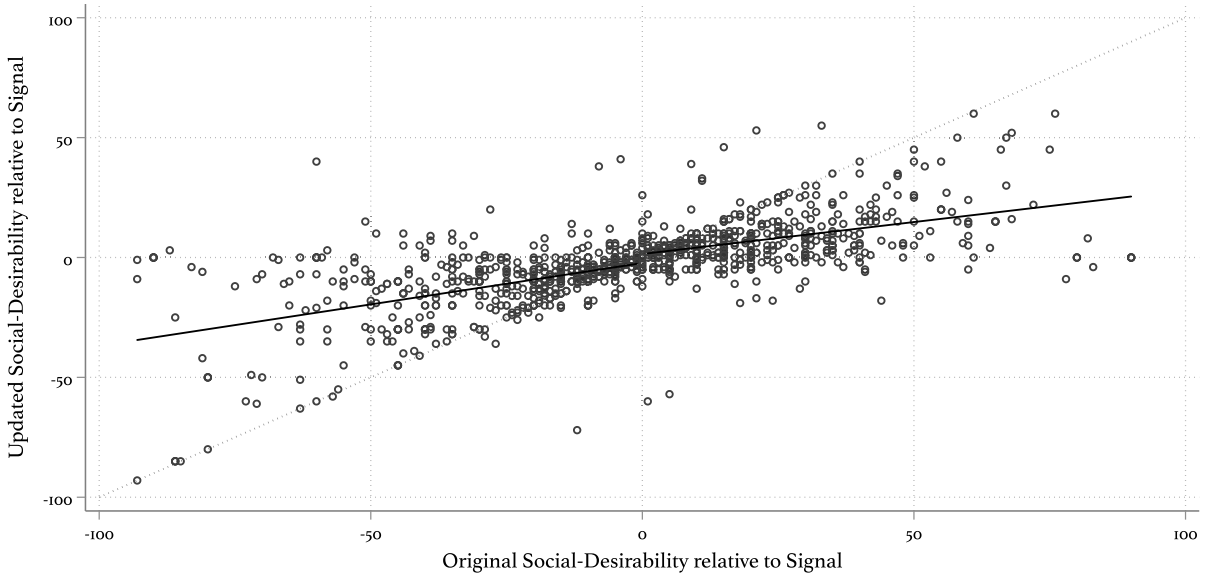

Fig. A.4. Predictors receiving signals from the H group.

A.5. Sentiment variance and updating

In this section, we construct a measure of polarization or uncertainty by calculating the variance in the sentiment scores associated with an action. Those with higher sentiment scores have less agreement among the sentiment group about whether the action is virtuous or stigmatized. In Table A.8, we regress this measure of variance on different measures of predictor confidence and accuracy in predictors' guesses.

A.6. Simulations

The following simulation results show a predictor's optimal Bayesian posterior guess conditional on the information source (IC signals versus H signals). Each plot shows the updated guess as a function of the signal under four different conditions: signals drawn

Table A.8
Sentiment variance and predictor accuracy and confidence.

	Initial Confidence	Updated Confidence	Initial Error (Absolute Value)
Variance of Sentiment	0.017***	-0.001	-0.171*
	(0.004)	(0.005)	(0.090)
Initial Error (Absolute)	0.003***	-0.007***	
	(0.001)	(0.001)	
Updated Error (Absolute)		0.003**	
		(0.002)	
Variance of Sentiment × IC Info Source		0.000	
		(0.006)	
Initial Confidence		0.485***	
		(0.023)	
IC Info Source		-0.041	
		(0.072)	
Constant	-0.174***	0.134***	22.643***
	(0.056)	(0.050)	(0.775)
Observations	2168		
Clusters	271		
Fixed-Effects:	None	None	None

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the subject level. All confidence values are standardized to have a mean of zero and standard deviation of 1. Action-level fixed effects are omitted because sentiment variance does not change within an action. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

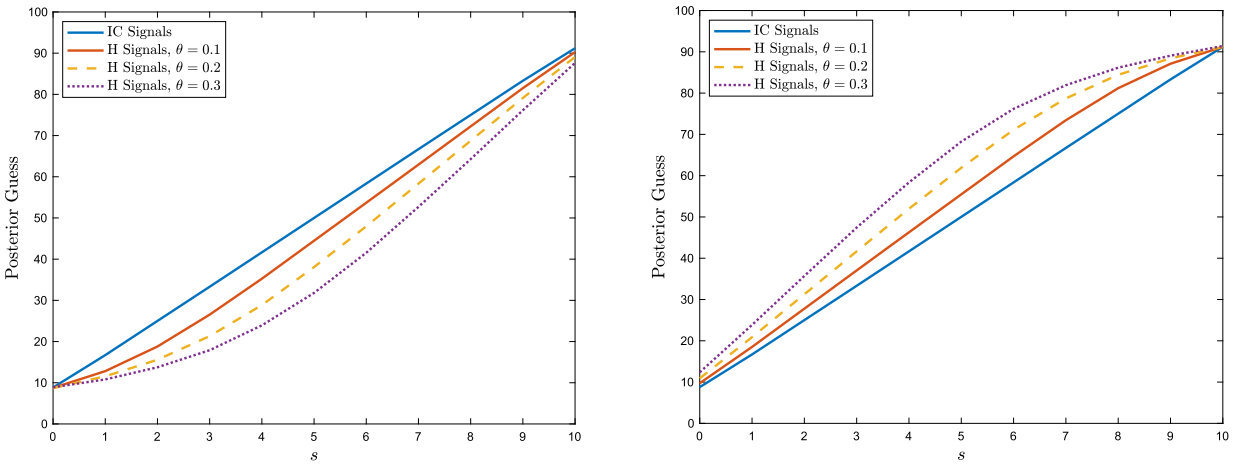

Fig. A.5. Bayesian predictions given the observed signal and information source for various values of θ and a uniform prior over IC. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

from the IC group, and signals drawn from the H group for each $\theta \in \{0.1, 0.2, 0.3\}$.[41] We construct such a plot for each of two cases: (a) the good under consideration is commonly viewed as virtuous and thus generates over-reporting, and (b) the good is commonly viewed as stigmatized and thus leads to under-reporting.

We must additionally specify the predictor's prior over the choice rate in the IC group, IC. We perform the exercise described above for three prior distributions: uniform (Fig. A.5), Beta(2,4) (Fig. A.6) and Beta(4,2) (Fig. A.7). See Fig. A.10 for a depiction of the prior distributions used, and see the end of this subsection for more details on how the simulations were run.

Differences between updated guesses and initial guesses We can also use the simulations above to show how a fully sophisticated predictor should change their guess after observing the signal. The plots below (Figs. A.8 and A.9) show the difference between the

[41] For simplicity, we follow the simple model in Section 3 in which there is common agreement about the valence of the action.

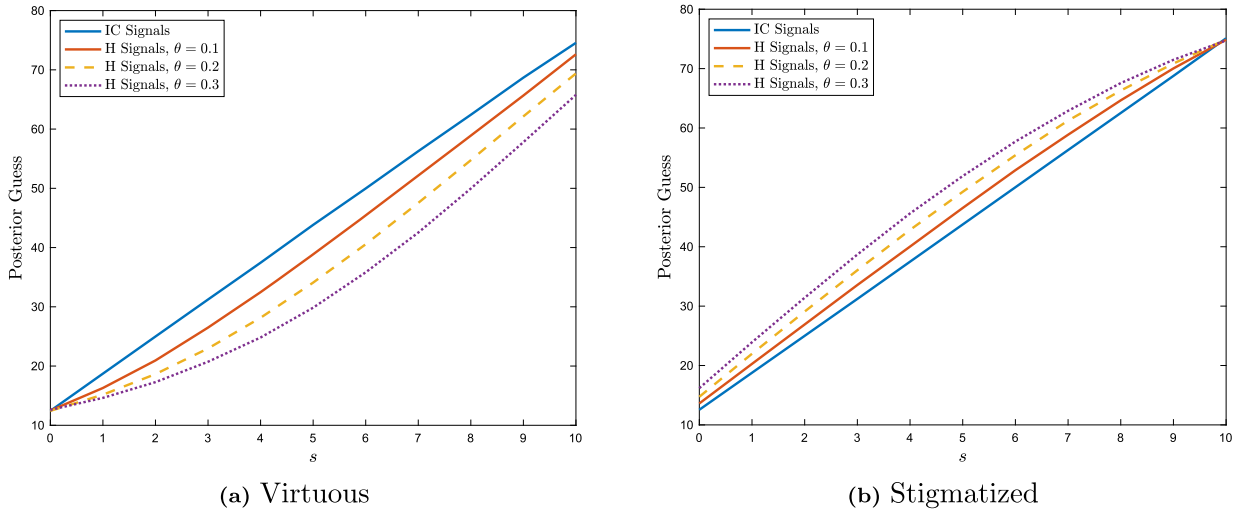

Fig. A.6. Bayesian predictions given the observed signal and information source for various values of θ and a Beta(2,4) prior over IC.

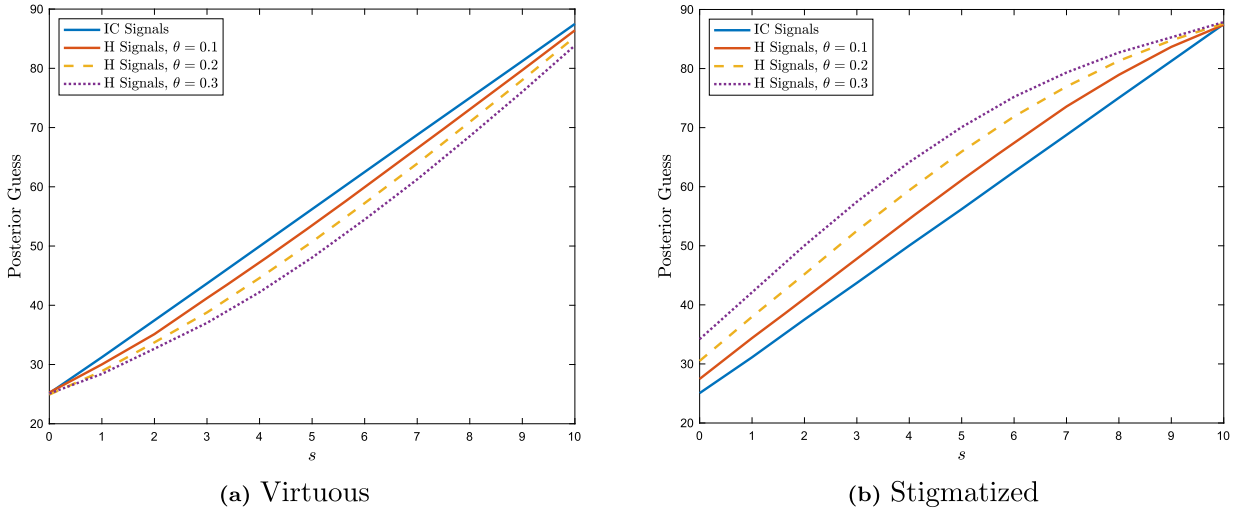

Fig. A.7. Bayesian predictions given the observed signal and information source for various values of θ and a Beta(4,2) prior over IC.

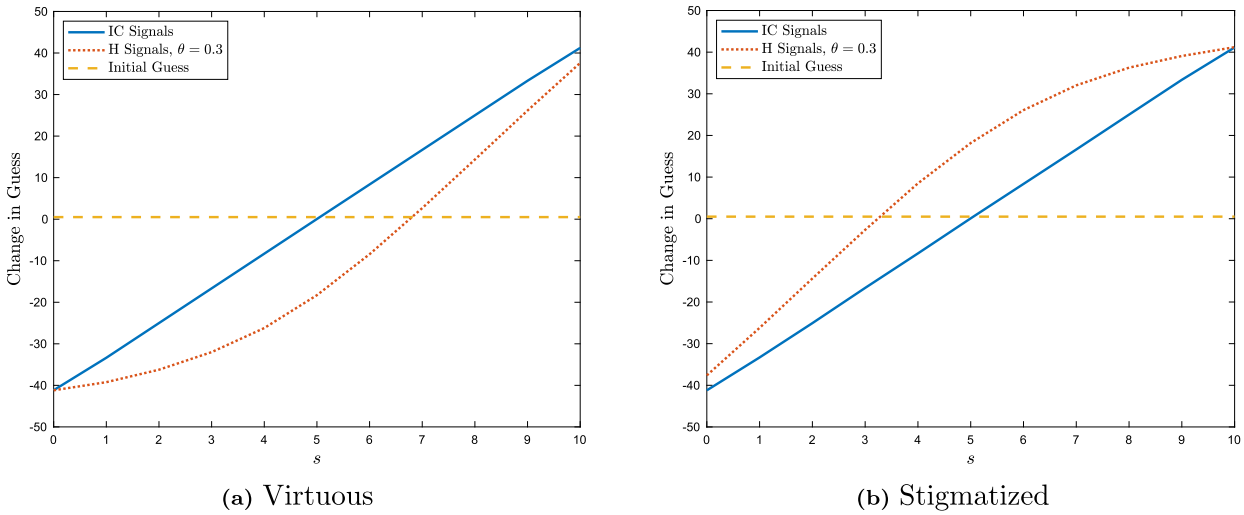

Fig. A.8. Difference between the second and first guess given the observed signal and information source for a uniform prior over IC.

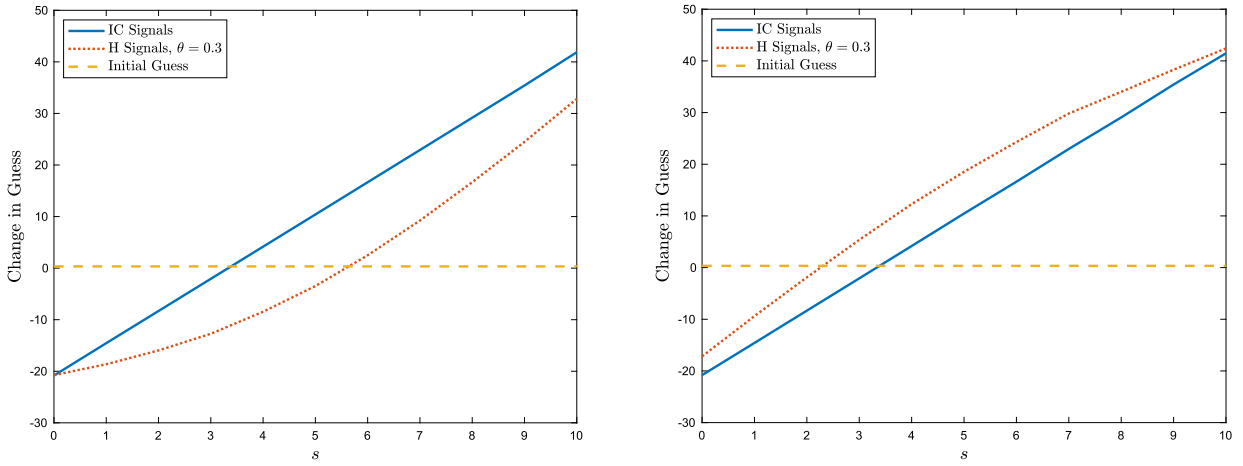

Fig. A.9. Difference between the second and first guess given the observed signal and information source for a Beta(2,4) prior over IC.

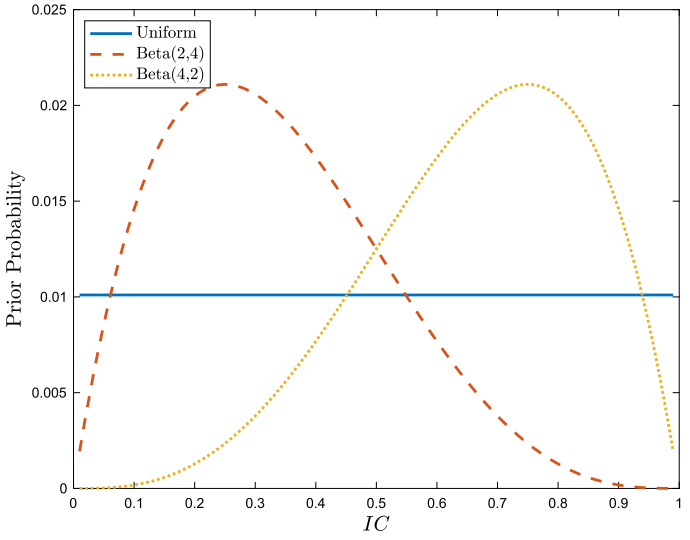

Fig. A.10. Prior distributions considered in the simulations.

second and first guess conditional on the signal and information source. For this exercise, we focus on $\theta = 0.3$, and consider two different priors: uniform and Beta(2,4).

Simulation details For each simulation, we consider a discrete space of possible values of the IC-group choice rate, IC. In particular, $IC \in \mathcal{I} \equiv \{0.01, 0.02, \ldots, 0.99\}$. First, we choose the predictor's prior over $\mathcal{I}$, denoted by p. We model this prior as a discrete approximation of a continuous Beta distribution by letting $p(IC) = f(IC)/\sum_{IC' \in \mathcal{I}} f(IC')$, where f is the continuous density under consideration.

Next, for each $IC \in \mathcal{I}$, we independently draw 10,000 IC-group signals and 10,000 H-group signals conditional on IC. Specifically, IC-group signals, denoted by s_{IC}, are random i.i.d. draws from Bin$(10, IC)$. We consider two cases for H-group signals, denoted by s_H, for each specified value of θ. In the "virtuous" case, s_H are i.i.d. draws from Bin$(10, \theta + (1-\theta)IC)$, and in the "stigmatized" case s_H are i.i.d. draws from Bin$(10, (1-\theta)IC)$. With our sets of s_{IC} and s_H signals in hand, we estimate the conditional probability of each signal realization conditional on IC. We denote this probability by $\hat{p}_{IC}(s|IC)$ for IV signals and $\hat{p}_H(s|IC)$ for H signals.

We can then construct the Bayesian predictor's updated beliefs over $\mathcal{I}$ conditional on the signal and information source. These beliefs for IC-signals and H-signals are, respectively, given by:

$$\hat{p}_{IC}(IC|s) = \frac{\hat{p}_{IC}(s|IC)p(IC)}{\sum_{IC' \in \mathcal{I}} \hat{p}_{IC}(s|IC')p(IC')}$$

and

$$\hat{p}_H(IC|s) = \frac{\hat{p}_H(s|IC)p(IC)}{\sum_{IC' \in \mathcal{I}} \hat{p}_H(s|IC')p(IC')}.$$

The predictor's updated expectation of IC given the signal is $\mathbb{E}_{IC}[IC|s] = \sum_{IC' \in \mathcal{I}} \hat{p}_{IC}(IC'|s) \cdot IC'$ for IC signals and $\mathbb{E}_H[IC|s] = \sum_{IC' \in \mathcal{I}} \hat{p}_H(IC'|s) \cdot IC'$ for H signals. The figures above plot these conditional expectations for various values of θ—$\theta \in \{0.1, 0.2, 0.3\}$—and different underlying priors—uniform, Beta(2,4), and Beta(4,2). We present the conditional expectations scaled by 100 so they are in percentage units.

Appendix B. Details on analyses

This appendix provides details about the specific regressions underlying our results in Sections 4–5. Our analysis carefully follows our pre-registration, which specifies an analysis of covariance (ANCOVA) framework. The sections below mirror the order of our results in Sections 4–5, and each indicates any changes to the analysis from the pre-registration along with any supplemental analyses that we conduct.

B.1. Manipulation check 1: SDR

In Column 1 of Table 3 we run the following pre-registered regression using each of the eight actions as an observation:

$$SDR_A = \beta_0 + \beta_1 \times V_A + \epsilon_A, \tag{B.1}$$

where V_A (defined in Equation (1)) is the average sentiment for action A across all participants in the Sentiment Study.

Alternative Specification: Individual-Level Sentiment

Our pre-registered analysis fails to take advantage of the full sample of subjects in the sentiment analysis. Thus, in Column 2 of Table 3, we include a supplementary analysis at the subject-level that increases statistical power without changing the underlying data. Following the standardized index defined in Equation (2), we generate $\widehat{V}_{i,A} \equiv \frac{V_{i,A} - \overline{V}_i}{\sigma_i}$ and include it on the right-hand side of the random-effects linear regression:

$$SDR_A = \beta_0 + \beta_1 \times \widehat{V}_{i,A} + \nu_i + \epsilon_{i,A}. \tag{B.2}$$

B.2. Manipulation check 2: accuracy

In Columns 1 and 3 of Table 4, we run pre-registered random-effects linear regressions to test for the impact of the signal source on the accuracy of the guesses:

$$ABS_{i,2,A} = \beta_0 + \beta_1 ABS_{i,1,A} + \beta_2 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \tag{B.3}$$

$$SQ_{i,2,A} = \beta_0 + \beta_1 SQ_{i,1,A} + \beta_2 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \tag{B.4}$$

where IC_i is an indicator variable equal to one if subject i received a signal from the IC group, and ν_i are subject random-effects (meaning they will not be individually identified). Standard errors are clustered at the individual level.

Alternative Specification: Individual Changes in Accuracy

Our pre-registered analysis takes the form of an analysis of covariance (ANCOVA). In Columns 2 and 4 of Table 4, we look at individual-level changes in accuracy to gain statistical power without changing the underlying data: $\Delta ABS_{i,A} = ABS_{i,2,A} - ABS_{i,1,A}$ and $\Delta SQ_{i,A} = SQ_{i,2,A} - SQ_{i,1,A}$. This is equivalent to restricting $\beta_1 = 1$ in our original equation. We repeat the random-effects linear regression with the new dependent variable:

$$\Delta ABS_{i,A} = \beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \tag{B.5}$$

$$\Delta SQ_{i,A} = \beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}. \tag{B.6}$$

B.3. Hypothesis 1: anticipation of SDR

In Column 1 of Table 5, we run the pre-registered random-effects linear regression:

$$\text{GUESS}_{i,2,A} = \beta_0 + \beta_1 \text{GUESS}_{i,1,A} + \beta_2 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \beta_5 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}, \tag{B.7}$$

where $S_{i,A}$ is the signal received by subject i for action A (i.e., the fraction of subjects from i's random sample of 10 who took action A), and $\bar{S}_{T,A}$ is the mean of the distribution of signals from group T (either IC or H) for action A. By controlling for $\bar{S}_{T,A}$, we are

able to use $S_{i,A}$ to identify the effect of a change in the signal that is derived only from sampling variation—that is, the mechanically-random change in the signal. δ_A are fixed-effects for each action. Again, ν_i are subject random-effects, and we cluster standard errors at the individual level.

Alternative Specification: Individual Changes

Alongside our pre-registered analysis, in Column 2 of Table 5, we include a higher-powered test of individual-level updating: $\Delta\text{GUESS}_{i,A} = \text{GUESS}_{i,2,A} - \text{GUESS}_{i,1,A}$. We also modified the specification to use fixed effects for all 16 combinations of actions and choice groups, $\delta_{T,A}$, rather than fixed-effects for actions and controls for signal means. Our alternate specification is:

$$\Delta\text{GUESS}_{i,A} = \beta_0 + \beta_1 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \delta_{T,A} + \nu_i + \epsilon_{i,A}. \tag{B.8}$$

Alternative Specification: Extensive- and Intensive-Margin Responses

Table A.5 provides an entirely new analysis of responses to signals. Column 1 estimates the probability of no response to the signal, Column 2 estimates the probability of exactly matching (i.e. perfectly responding to) the signal, and Column 3 explores intermediate responses where the predictor neither ignores nor matches the signal. The three estimating equations are included in sequence below:

$$\Pr\left(\text{MATCH}_{i,A}\right) = \Phi\left(\beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}\right), \tag{B.9}$$

$$\Pr\left(\text{IGNORE}_{i,A}\right) = \Phi\left(\beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}\right), \tag{B.10}$$

$$\Delta\text{GUESS}_{i,A} = \beta_0 + \beta_1 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \delta_{T,A} + \nu_i + \epsilon_{i,A}, \tag{B.11}$$

where $\text{MATCH}_{i,A}$ and $\text{IGNORE}_{i,A}$ are indicators for $\text{GUESS}_{2,A} = S_{i,A}$ and $\text{GUESS}_{2,A} = \text{GUESS}_{1,A}$, respectively. The third equation is estimated on a selected sample of guesses that excludes any where $\text{MATCH}_{i,A} = 1$ or $\text{IGNORE}_{i,A} = 1$.

B.4. Hypothesis 2: direction of SDR

Hypothesis 2 was not included in our pre-registration. All results can be found in Table 6.

To test this hypothesis, we must divide our sample based on whether or not the signal is SDR-congruent—that is, if it suggests that the observed behavior is more or less socially desirable than the predictor's initial guess. The direction of social desirability will be determined based on the relative selection rates for the full sample. An action is socially desirable if $SDR_A > 0$; thus, one approach would be to classify a signal as SDR-congruent if it suggests that there are more people engaging in (or claiming to engage in) this action than the predictor initially guessed. The opposite would be true for actions that are socially undesirable (i.e. $SDR_A < 0$). Unfortunately, a predictor's initial guess is endogenous, so we must instrument for their initial guess to avoid this problem. We, therefore, consider a signal to be SDR-congruent if it is above the mean initial guess in the population for socially desirable actions or is below the mean initial guess in the population for socially undesirable actions.

Column 1 of Table 6 presents our first test of Hypothesis 2 using the same random-effects linear-regression specification as in Equation (B.7), but including a full set of interactions with terms that indicate whether the signal is SDR-congruent or SDR-dissonant:

$$\begin{aligned}\text{GUESS}_{i,2,A} = \beta_0 + \beta_1 \text{GUESS}_{i,1,A} + \beta_2 S_{i,A} \times \text{PI}_{i,A} \\ + \beta_3 S_{i,A} \times \text{IC}_i \times \text{PI}_{i,A} + \beta_4 \text{IC}_i \times \text{PI}_{i,A} + \beta_5 \text{PI}_{i,A} + \beta_6 S_{i,A} \times \text{PD}_{i,A} \\ + \beta_7 S_{i,A} \times \text{IC}_i \times \text{PD}_{i,A} + \beta_8 \text{IC}_i \times \text{PD}_{i,A} + \beta_9 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}.\end{aligned} \tag{B.12}$$

Here, we interact all of the relevant terms from Equation (B.7) with $\text{PI}_{i,A}$ ($\text{PD}_{i,A}$), indicators for whether the signal is SDR-congruent (SDR-dissonant) relative to $\text{GUESS}_{i,1,A}$. We test two aspects of updating: (1) if signals from the IC group are weighted more heavily (relative to signals from the H group) as they indicate greater image inflation (i.e. if $\beta_3 > 0$) and (2) if signals from the H group are weighted more heavily (relative to signals from the IC group) as they indicate image deflation (i.e. if $\beta_7 < 0$).

Alternative Specification: Individual Changes

Similar to Equation (B.8), we include our measure of individual-level updating, $\Delta\text{GUESS}_{i,A}$, and our fixed-effects for combinations of action and group, $\delta_{T,A}$, in place of $\bar{S}_{T,A}$ and δ_A. This analysis is presented in Column 2 of Table 6.

B.5. Hypothesis 3: relative magnitude of SDR

Column 1 of Table 7 conducts our pre-registered test of Hypothesis 3 using the same random-effects linear-regression specification as in Equation (B.7). However, we now include terms interacted with the absolute value of our measure of SDR:

$$\text{Guess}_{i,2,A} = \beta_0 + \beta_1 \text{Guess}_{i,1,A} + \beta_2 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \beta_5 S_{i,A} \times |\text{SDR}_A| + \beta_6 S_{i,A} \times \text{IC}_i \times |\text{SDR}_A| + \beta_7 \text{IC}_i \times |\text{SDR}_A| + \beta_8 |\text{SDR}_A| + \beta_9 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}. \quad \text{(B.13)}$$

Here, we interact all of the relevant terms from Equation (B.7) with the absolute value of our measure of SDR for action A, $|SDR_A|$. We test if signals from the IC group are weighted more heavily (relative to signals from the H group) as SDR becomes more extreme (i.e. if $\beta_6 > 0$).

Alternative Specification: Individual Changes and Sensitivity to Sentiment

Similar to Equation (B.8), we include our measure of individual-level updating, $\Delta\text{Guess}_{i,A}$, and our fixed-effects for combinations of action and group, $\delta_{T,A}$, in place of $\bar{S}_{T,A}$ and δ_A. This analysis is presented in Column 2 of Table 7

We also measure how sensitive subjects are to changes in our proxy for social desirability, sentiment. Specifically, we replace $|\text{SDR}_A|$ with a standardized measure of how extreme sentiment is toward the action, $|\widehat{V}_A| = \frac{|V_A - \overline{V}|}{\sigma_V}$, where $\overline{V}$ and σ_V are the mean and standard deviation of V_A across all eight actions. This analysis is presented in Column 3 of Table 7.

B.6. Confidence

Our exploratory analysis on confidence was not pre-registered but adds substantively to our understanding of the implications of poor data-quality on behaviors surrounding inference (in this case, confidence in guesses). Table 8 presents two tests of the impact of a predictor's accuracy on their confidence. Prior to running this analysis, we normalize confidence measures across all predictors and all actions to generate $\widehat{\text{Confidence}}_{i,1,A}$ and $\widehat{\text{Confidence}}_{i,2,A}$, both of which have mean 0 and standard deviation 1. Column 1 presents the association between normalized confidence and the accuracy of initial guesses using the following specification:

$$\widehat{\text{Confidence}}_{i,1,A} = \beta_0 + \beta_1 ABS_{i,1,A} + \delta_A + \nu_i + \epsilon_{i,A}, \quad \text{(B.14)}$$

where $ABS_{i,1,A}$ is the absolute error in subject i's initial guess, δ_A is a vector of action fixed effects, and ν_i are subject random-effects. Standard errors are clustered at the individual level.

Column 2 of Table 8 demonstrates how this confidence evolves after receiving information. It uses the following specification:

$$\widehat{\text{Confidence}}_{i,2,A} = \beta_0 + \beta_1 ABS_{i,1,A} + \beta_2 ABS_{i,2,A} + \beta_3 ABS_{i,2,A} \times \text{IC}_i + \beta_4 \widehat{\text{Confidence}}_{i,1,A} + \beta_5 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \quad \text{(B.15)}$$

where $ABS_{i,2,A}$ is the absolute error in subject i's updated guess. Standard errors are again clustered at the individual level.

In both tests, we consider how confidence is associated with accuracy (β_1 in Equation (B.14) and β_2 in Equation (B.15)). In Equation (B.15), we also care about how this depends on the randomly-assigned information source (β_3).

B.7. Experience with SDR

Column 1 of Table A.6 presents the pre-registered test of our hypothesis about experience. We use the same random-effects linear-regression specification as in Equation (B.7), but include terms interacted with the role that Predictor i played in the Choice Study:

$$\text{Guess}_{i,2,A} = \beta_0 + \beta_1 \text{Guess}_{i,1,A} + \beta_2 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \beta_5 S_{i,A} \times \text{Exp}_i + \beta_6 S_{i,A} \times \text{IC}_i \times \text{Exp}_i + \beta_7 \text{IC}_i \times \text{Exp}_i + \beta_8 \text{Exp}_i + \beta_9 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}, \quad \text{(B.16)}$$

where Exp_i is an indicator variable equal to one if the predictor has previous experience participating in the IC or H group. Again, we test for a significant interaction effect by testing if $\beta_6 > 0$.

We repeat this analysis looking at members of the H and IC groups separately, which reveals heterogeneity in the learned experience of the two groups.

Columns 3 and 4 of Table 5 adapt this approach by isolating the updating behavior of new predictors and experienced predictors, respectively.

Alternative Specification: Individual Changes

As with Hypotheses 1–3, we replicate the pre-registered analysis with an alternative specification. As before, we include individual-level updating and fixed-effects for combinations of action and group. This analysis is presented in Column 2 of Table A.6.

Appendix C. Experimental instructions

C.1. Sentiment-study instructions

Thank you for your participation today. Just for participating in this study, you will receive $5 toward your Take-Home Pay. In order to receive your Take-Home Pay, you need to complete the entire survey and then instructions for payment will be emailed to you once all responses have been collected.

All of the choices will be made in private. This means that your responses will be observed by the researchers after-the-fact and no one else.

This is a non-deceptive experiment. That means that, if we say an action has real consequences, those consequences will actually happen. On the other hand, if a choice is hypothetical, we will tell you in advance that it is hypothetical.

C.1.1. Sentiment-study comprehension question

We will be asking you to respond to questions about a series of potential scenarios. Your responses will not have any real consequences, we are simply asking for your feelings on each scenario.

To ensure that you understand, please answer the following question. Will your choices have real consequences?

- Yes, all of them will count.
- Yes, on will be chosen at randomly-chosen.
- No, you are just asking my opinion.

C.2. Choice-study instructions: hypothetical group

Thank you for your participation today. Just for participating in this part of the experiment, you will receive $5 toward your Take-Home Pay. In order to receive your Take-Home Pay, you must complete the second part of the experiment that we will email to you after you complete this. The second part of the experiment will pay you between $5 and $10. So, you will receive between $10 and $15 for completing both parts of the study.

All of the choices will be made in private. This means that your choice will be observed by the researchers after-the-fact and no one else.

This is a non-deceptive experiment. That means that, if we say an action has real consequences, those consequences will actually happen. On the other hand, if a choice is hypothetical, we will tell you in advance that it is hypothetical.

C.2.1. Choice-study comprehension question: hypothetical

We will be asking you to make a series of choices and answer a few questions. All of your choices will be hypothetical. Meaning that none of your choices will have real consequences.

We simply want to know how you would respond if you were asked to make a choice in these hypothetical situations.

To ensure that you understand, please answer the following question. Will your choices have real consequences?

- Yes, one randomly selected choice will count
- Yes, all of them will count.
- No, they are hypothetical.

C.3. Choice-study instructions: incentive compatible group

Thank you for your participation today. Just for participating in this part of the experiment, you will receive $5 toward your Take-Home Pay. In order to receive your Take-Home Pay, you must complete the second part of the experiment that we will email to you after you complete this. The second part of the experiment will pay you between $5 and $10. So, you will receive between $10 and $15 for completing both parts of the study.

All of the choices will be made in private. This means that your choice will be observed by the researchers after-the-fact and no one else.

This is a non-deceptive experiment. That means that, if we say an action has real consequences, those consequences will actually happen. On the other hand, if a choice is hypothetical, we will tell you in advance that it is hypothetical.

C.3.1. Choice-study comprehension question: incentive-compatible

We will be asking you to make a series of choices and answer a few questions. Your choices will have real consequences.

At the end of the study, we will randomly select one of your choices to be the Choice That Counts. The Choice That Counts will determine your outcome today. Since any choice can be selected as the Choice That Counts, you should treat every choice like it is the Choice That Counts.

To reiterate, only one of your choices will be randomly chosen as the Choice That Counts. So, treat each choice as a separate, meaningful choice.

To ensure that you understand, please answer the following question. Will your choices have real consequences?

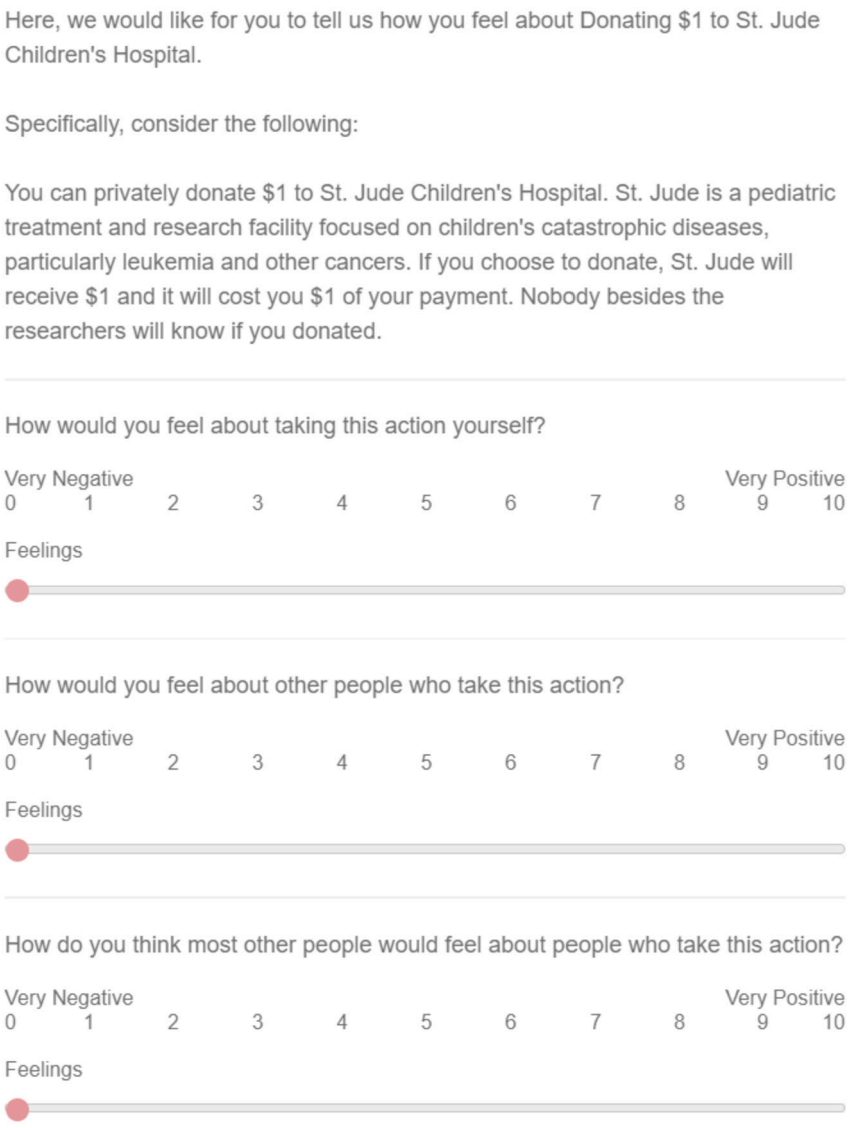

Fig. C.1. Sentiment-Study Decision Screen.

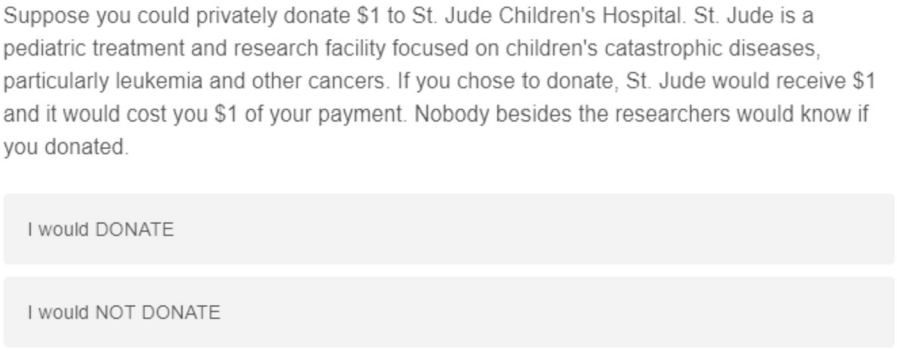

Fig. C.2. Hypothetical Decision.

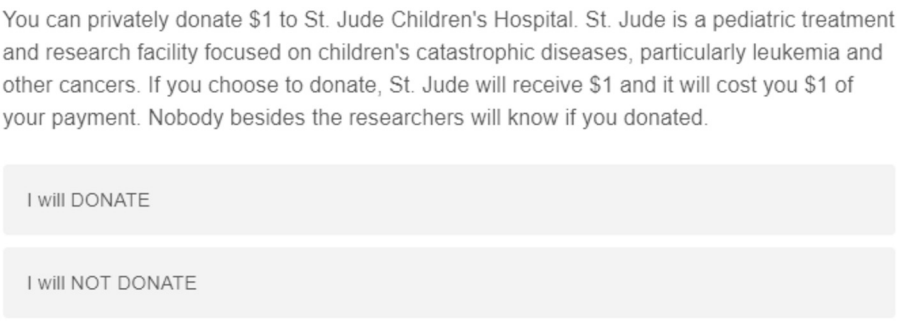

Fig. C.3. IC Decision.

- Yes, on randomly selected choice will count
- Yes, all of them will count.
- No, they are hypothetical.

C.4. Prediction-study instructions

C.4.1. Prediction-study general instructions

Just for participating, you will be guaranteed to receive $5. You may earn significantly more money depending on how you perform your tasks in this study.

In this study, you are a "Predictor." Your task today will be to make predictions about the behavior of other participants in the study. The more accurate your predictions are, the more money you will earn.

We recruited students at the University of Arkansas to be "Real-Deciders." Real-Deciders made a series of private choices and entered them confidentially into a computer.

The Real-Deciders knew that their choices would never be individually observed by anyone but the researchers.

The choices that the Real-Deciders made had real consequences. One choice made by each Real-Decider was randomly selected to be carried out by the experimenters.

Key Points: Real-Deciders made private decisions without anyone watching. Their decisions had real consequences and really determined their payment.

C.4.2. Prediction-study comprehension question

What is your role in this study?

- Make decisions
- Guess what decisions the Real-Deciders made
- Help the Real-Deciders make their decisions

Did the Real-Decides' choices have consequences?

- Yes, their choices mattered
- No, their choices were hypothetical

C.4.3. Prediction-study predictions instructions

The Real-Deciders made decisions about several different actions. We described these actions to the Real-Deciders before they made their choices. We will describe them to you in exactly the same way.

For each action, there were only two options: Option 1: Take the action Option 2: Do not take the action

Your job is to predict **P** – the number of the Real-Deciders out of 100 who chose to take the action (the first option). You will report your best guess about **P**.

There is a true percentage of Real-Deciders who chose to take each action. We'll call this value **"True-P"**. The closer you get to guessing the **True-P**, the more money you can earn.

It is important that you think carefully about your prediction for **P** because we will offer you a chance to win money based on your accuracy.

You will make 16 predictions in this study. We will randomly select one of these predictions to be the Prediction That Counts. Your money will depend on how accurate you are on the Prediction That Counts. Since each prediction could be the Prediction That Counts, you should treat each prediction like it is the Prediction That Counts.

C.4.4. Payment comprehension question

How many of your 16 predictions will determine your payment?

- All of them collectively
- One selected at random: "the Prediction That Counts"
- The first one
- The last one

C.4.5. Prediction-study lottery draw instructions

You will have a chance to earn an extra $5 lottery bonus at the end of the study (in addition to the $5 you are already guaranteed). You will earn lottery tickets if your guess about **P** is close to the **True-P**. At the end of the session, we will randomly draw a lottery number between 1 and 100; if that number matches one of your lottery tickets, you will win the bonus payment. So it's best to get as many lottery tickets as possible to maximize your chance of a bonus.

On the next page, we will describe how you can earn tickets based on your guess of **P**. The precise method we use to calculate your lottery tickets may sound complicated, but you will always earn the most if you simply answer truthfully.

C.4.6. Prediction-study lottery draw comprehension questions

What is the easiest way to earn the most lottery tickets?

- Guess the largest number as the **True-P**
- Guess the smallest number as the **True-P**
- Guess your honest beliefs about the **True-P**

C.4.7. Prediction-study lottery ticket instructions

The number of lottery tickets you will receive will be one of the following: ***Option A***: The number of lottery tickets you will receive is equal to the **True-P**. ***Option B***: The number of lottery tickets you will receive is equal to your "Random Draw," which is a random number between 0 and 100.

The option you receive depends on how your Random Draw compares to your guess about **P**. If your Random Draw is below your guess, then you will get Option A (lottery tickets equal to the **True-P**). If your Random Draw is above your guess, then you will get Option B (lottery tickets equal to your Random Draw).

Here are two examples:

If your guess is that **P = 50**, and your Random Draw is 25, then your Random Draw is less than your guess about the **True-P**. So, you will get Option A (lottery tickets equal to the **True-P**).

If your guess is that **P = 50**, and your Random Draw is 75, then your Random Draw is more than your guess about the **True-P**. So, you will get Option B (lottery tickets equal to your Random Draw).

C.4.8. Prediction-study lottery ticket comprehension questions

If your guess about **P** is that **P = 23** and your Random Draw is 17, how many lottery tickets will you receive?

- 50
- Option A: you will receive a number of lottery tickets equal to the **True-P**
- Option B: you will receive a number of lottery tickets equal to your Random Draw, 17.

If your guess about **P** is that **P = 43** and your Random Draw is 73, how many lottery tickets will you receive?

- 50
- Option A: you will receive a number of lottery tickets equal to the **True-P**
- Option B: you will receive a number of lottery tickets equal to your Random Draw, 73.

You might think you can "game the system" and earn more lottery tickets by reporting a higher guess for P than you really believe. That won't help you. It will only increase the chance that you pass up your Random Draw when it is a high number.

On the other hand, you also can't game the system by reporting a lower guess for P than you really believe. If you do that, then you will increase the chance that you accept your Random Draw when it is a low number.

C.4.9. 2nd prediction instructions (hypothetical information)

Your task is to predict the behavior of the 100 Real-Deciders that we recruited from the University of Arkansas to participate in the study. Before you make these predictions for a second time, we will show you the decisions of 10 "Hypothetical-Deciders."

We recruited 100 Hypothetical-Deciders at the same time that we recruited the 100 Real-Deciders for the study. Both were recruited out of the same subject pool at the University of Arkansas.

For every one of the decisions that the Real-Deciders made, the Hypothetical-Deciders reported what they would have chosen if they had been asked to choose. But, the statements made by Hypothetical-Deciders did not have any real consequences.

If a Hypothetical-Decider reported that they would take an action, the Hypothetical-Deciders never actually had to take the action. These responses were entirely hypothetical.

We have randomly selected 10 of the 100 Hypothetical-Deciders. We will show you their responses on all 8 actions.

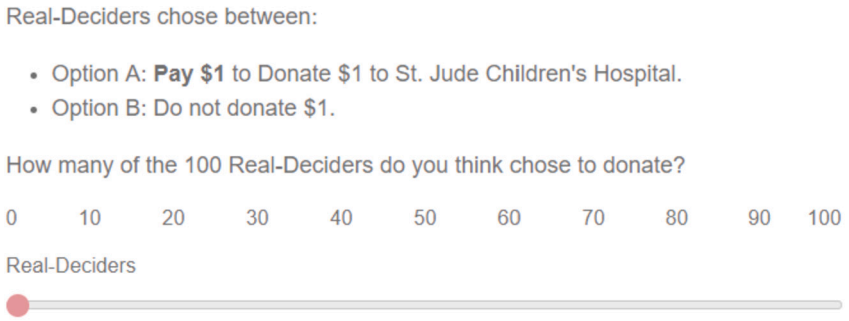

Fig. C.4. First-Prediction Choice.

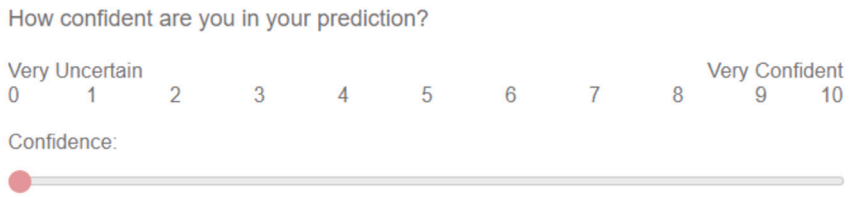

Fig. C.5. First-Prediction Choice Confidence.

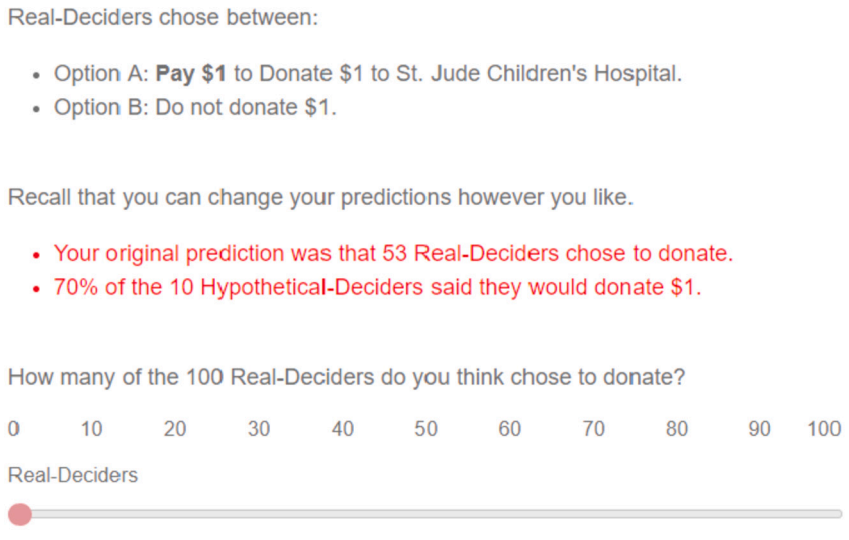

Fig. C.6. Second Prediction Choice.

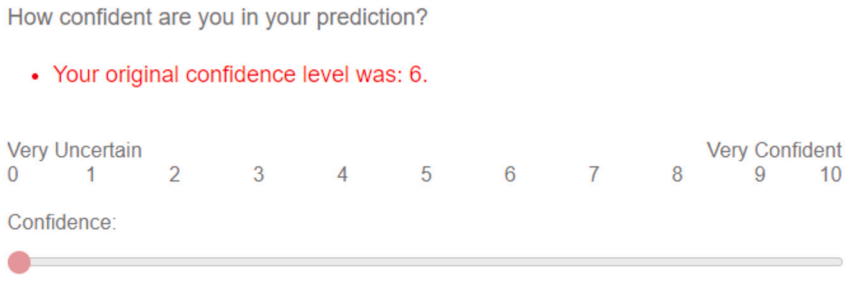

Fig. C.7. Second Prediction Choice Confidence.

The Hypothetical-Deciders did not make the exact same choices as the Real-Deciders. But this information may be useful in revising your predictions about the choices that the 100 Real-Deciders made.

While you are revising your predictions about the Real-Deciders, we will remind you of the responses of the Hypothetical-Deciders. So, you do not need to memorize their choices now.

C.4.10. 2nd prediction comprehension question (hypothetical information)

Did the Hypothetical-Deciders make choices with actual consequences?

- Yes, their choices mattered
- No, their choices were hypothetical

C.4.11. 2nd prediction instructions (IC information)

Your task is to predict the behavior of the 100 Real-Deciders that we recruited from the University of Arkansas to participate in the study. Before you make these predictions for a second time, we will show you the decisions of 10 of the Real-Deciders.

These 10 Real-Deciders were randomly selected from among the 100 Real-Deciders you are making predictions about. They were all recruited from the same subject pool at the University of Arkansas.

Recall that all choices made by the Real-Deciders had real consequences.

We have randomly selected 10 of the 100 Real-Deciders. We will show you their choices on all 8 actions.

The 10 randomly chosen Real-Deciders that we will show you did not make the exact same choices as the other 90 Real-Deciders. But this information may be useful in revising your predictions about the choices that all 100 Real-Deciders made.

While you are revising your predictions about the 100 Real-Deciders, we will remind you of the responses of the 10 randomly chosen Real-Deciders. So, you do not need to memorize their choices now.

C.4.12. 2nd prediction comprehension question (IC information)

Did the 10 randomly selected Real-Deciders make choices with actual consequences?

- Yes, their choices mattered
- No, their choices were hypothetical

Data availability

Data will be made available on request.

References

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preferences for truth-telling. Econometrica 87, 1115–1153.
Adkins, R.E., Dowdle, A.J., 2002. The money primary: what influences the outcome of pre-primary presidential nomination fundraising? Pres. Stud. Q. 32, 256–275.
Agranov, M., Goeree, J.K., Romero, J., Yariv, L., 2018. What makes voters turn out: the effects of polls and beliefs. J. Eur. Econ. Assoc. 16, 825–856.
Araujo, F.A., Wang, S.W., Wilson, A.J., 2021. The times they are A-changing: experimenting with dynamic adverse selection. Am. Econ. J. Microecon. 13, 1–22.
Arnold, H.J., Feldman, D.C., Purbhoo, M., 1985. The role of social-desirability response bias in turnover research. Acad. Manag. J. 28, 955–966.
Augenblick, N., Rabin, M., 2019. An experiment on time preference and misprediction in unpleasant tasks. Rev. Econ. Stud. 86, 941–975.
Bagwell, L.S., Bernheim, B.D., 1996. Veblen effects in a theory of conspicuous consumption. Am. Econ. Rev., 349–373.
Becker, G.M., DeGroot, M.H., Marschak, J., 1964. Measuring utility by a single-response sequential method. Behav. Sci. 9, 226–232.
Bénabou, R., Tirole, J., 2002. Self-confidence and personal motivation. Q. J. Econ. 117, 871–915.
Benjamin, D.J., 2019. Errors in Probabilistic Reasoning and Judgment Biases. Handbook of Behavioral Economics: Applications and Foundations, vol. 2. Elsevier, pp. 69–186.
Bharadwaj, P., Pai, M.M., Suziedelyte, A., 2017. Mental health stigma. Econ. Lett. 159, 57–60.
Boukouras, A., Jennings, W., Li, L., Maniadis, Z., 2023. Can biased polls distort electoral results? Evidence from the lab. Eur. J. Polit. Econ. 78, 102383.
Braghieri, L. Political correctness, social image, and information transmission. American Economic Review. In press.
Brown-Iannuzzi, J.L., Najle, M.B., Gervais, W.M., 2019. The illusion of political tolerance: social desirability and self-reported voting preferences. Soc. Psychol. Pers. Sci. 10, 364–373.
Brownback, A., Novotny, A., 2018. Social desirability bias and polling errors in the 2016 presidential election. J. Behav. Exp. Econ. 74, 38–56.
Bursztyn, L., Cantoni, D., Funk, P., Schönenberger, F., Yuchtman, N., 2024. Identifying the effect of election closeness on voter turnout: evidence from Swiss referenda. J. Eur. Econ. Assoc. 22, 876–914.
Bursztyn, L., Ferman, B., Fiorin, S., Kanz, M., Rao, G., 2018. Status goods: experimental evidence from platinum credit cards. Q. J. Econ. 133, 1561–1595.
Cantú, F., Márquez, J., 2021. The effects of election polls in Mexico's 2018 presidential campaign. Elect. Stud. 73, 102379.
Carlson, E., 2018. The perils of pre-election polling: election cycles and the exacerbation of measurement error in illiberal regimes. Res. Polit. 5, 2053168018774728.
Charness, G., Oprea, R., Yuksel, S., 2021. How do people choose between biased information sources? Evidence from a laboratory experiment. J. Eur. Econ. Assoc. 19, 1656–1691.
Coffman, K.B., Coffman, L.C., Ericson, K.M.M., 2017. The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated. Manag. Sci. 63, 3168–3186.
Crawford, V., 1998. A survey of experiments on communication via cheap talk. J. Econ. Theory 78, 286–298.
Crawford, V.P., Sobel, J., 1982. Strategic information transmission. Econometrica, 1431–1451.
de Quidt, J., Haushofer, J., Roth, C., 2018. Measuring and bounding experimenter demand. Am. Econ. Rev. 108, 3266–3302.
Del Boca, F.K., Noll, J.A., 2000. Truth or consequences: the validity of self-report data in health services research on addictions. Addiction 95, 347–360.
Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G., 2011. Individual risk attitudes: measurement, determinants, and behavioral consequences. J. Eur. Econ. Assoc. 9, 522–550.

Edwards, A.L., 1957. The Social Desirability Variable in Personality Assessment and Research. Dryden Press.
Ericson, K.M.M., 2011. Forgetting we forget: overconfidence and memory. J. Eur. Econ. Assoc. 9, 43–60.
Esponda, I., Vespa, E., 2014. Hypothetical thinking and information extraction in the laboratory. Am. Econ. J. Microecon. 6, 180–202.
Farrell, J., Rabin, M., 1996. Cheap talk. J. Econ. Perspect. 10, 103–118.
Fedyk, A., 2018. Asymmetric naivete: Beliefs about self-control. Available at SSRN 2727499.
Finkel, S.E., Guterbock, T.M., Borg, M.J., 1991. Race-of-interviewer effects in a preelection poll Virginia 1989. Public Opin. Q. 55, 313–330.
Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise—an experimental study on cheating. J. Eur. Econ. Assoc. 11, 525–547.
Fox News, 2016. See which candidates qualified for the Fox News-Google GOP debates. Fox News. http://insider.foxnews.com/2016/01/26/lineup-republican-candidates-fox-news-google-gop-debates.
Gonzalez-Ocantos, E., De Jonge, C.K., Meléndez, C., Osorio, J., Nickerson, D.W., 2012. Vote buying and social desirability bias: experimental evidence from Nicaragua. Am. J. Polit. Sci. 56, 202–217.
Großer, J., Schram, A., 2010. Public opinion polls, voter turnout, and welfare: an experimental study. Am. J. Polit. Sci. 54, 700–717.
Haaland, I., Roth, C., Wohlfart, J., 2020. Designing information provision experiments. Working Paper 20/20, CEBI Working Paper Series.
Heerwig, J.A., McCabe, B.J., 2009. Education and social desirability bias: the case of a Black presidential candidate. Soc. Sci. Q. 90, 674–686.
Heidhues, P., Kőszegi, B., 2010. Exploiting naivete about self-control in the credit market. Am. Econ. Rev. 100, 2279–2303.
Holbrook, A.L., Krosnick, J.A., 2010. Social desirability bias in voter turnout reports: tests using the item count technique. Public Opin. Q. 74, 37–67.
Hopkins, D.J., 2009. No more wilder effect, never a Whitman effect: when and why polls mislead about Black and feMale candidates. J. Polit. 71, 769–781.
Janus, A.L., 2010. The influence of social desirability pressures on expressed immigration attitudes. Soc. Sci. Q. 91, 928–946.
Jones, A.E., Elliot, M., 2016. Examining social desirability in measures of religion and spirituality using the bogus pipeline. Rev. Relig. Res., 1–18.
Kahan, D.M., 2015. The politically motivated reasoning paradigm, part 1: what politically motivated reasoning is and how to measure it. In: Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource, pp. 1–16.
Kane, J.G., Craig, S.C., Wald, K.D., 2004. Religion and presidential politics in Florida: a list experiment. Soc. Sci. Q. 85, 281–293.
Karlan, D.S., Zinman, J., 2012. List randomization for sensitive behavior: an application for measuring use of loan proceeds. J. Dev. Econ. 98, 71–75.
Kartik, N., 2009. Strategic communication with lying costs. Rev. Econ. Stud. 76, 1359–1395.
Kruger, J., Dunning, D., 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J. Pers. Soc. Psychol. 77, 1121.
Krupka, E.L., Weber, R.A., 2013. Identifying social norms using coordination games: why does dictator game sharing vary? J. Eur. Econ. Assoc. 11, 495–524.
Krysan, M., 1998. Privacy and the expression of white racial attitudes: a comparison across three contexts. Public Opin. Q., 506–544.
Latkin, C.A., Edwards, C., Davey-Rothwell, M.A., Tobin, K.E., 2017. The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. Addict. Behav. 73, 133–136.
Maccoby, E.E., Maccoby, N., 1954. The interview: a tool of social science. In: Handbook of Social Psychology, vol. 1, pp. 449–487.
Martínez-Marquina, A., Niederle, M., Vespa, E., 2019. Failures in contingent reasoning: the role of uncertainty. Am. Econ. Rev. 109, 3437–3474.
Ngangoué, M.K., Weizsäcker, G., 2021. Learning from unrealized versus realized prices. Am. Econ. J. Microecon. 13, 174–201.
Paulhus, D.L., 1984. Two-component models of socially desirable responding. J. Pers. Soc. Psychol. 46, 598–609.
Powell, R.J., 2013. Social desirability bias in polling on same-sex marriage ballot measures. Am. Polit. Res. 41, 1052–1070.
Pronin, E., Lin, D.Y., Ross, L., 2002. The bias blind spot: perceptions of bias in self versus others. Pers. Soc. Psychol. Bull. 28, 369–381.
Raghavarao, D., Federer, W.T., 1979. Block total response as an alterNative to the randomized response method in surveys. J. R. Stat. Soc., Ser. B, Methodol., 40–45.
Reeves, K., et al., 1997. Voting Hopes or Fears?: White Voters, Black Candidates & Racial Politics in America. Oxford University Press on Demand.
Rosenfeld, B., Imai, K., Shapiro, J.N., 2016. An empirical validation study of popular survey methodologies for sensitive questions. Am. J. Polit. Sci. 60, 783–802.
Serra-Garcia, M., Gneezy, U., 2021. Mistakes, overconfidence, and the effect of sharing on detecting lies. Am. Econ. Rev. 111, 3160–3183.
Stephens-Davidowitz, S., 2014. The cost of racial animus on a black candidate: evidence using Google search data. J. Public Econ. 118, 26–40.
Streb, M.J., Burrell, B., Frederick, B., Genovese, M.A., 2008. Social desirability effects and support for a female American president. Public Opin. Q. 72, 76–89.
Thaler, M., 2024. The fake news effect: experimentally identifying motivated reasoning using trust in news. Am. Econ. J. Microecon. 16, 1–38.
Tourangeau, R., Rips, L.J., Rasinski, K., 2000. The Psychology of Survey Response. Cambridge University Press.
Tourangeau, R., Yan, T., 2007. Sensitive questions in surveys. Psychol. Bull. 133, 859–883.
Veblen, T., 1899. The Theory of the Leisure Class: An Economic Study of Institutions.
Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60, 63–69.
West, R.F., Meserve, R.J., Stanovich, K.E., 2012. Cognitive sophistication does not attenuate the bias blind spot. J. Pers. Soc. Psychol. 103, 506–519.
Yeargain, T., 2020. Fake polls, real consequences: the rise of fake polls and the case for criminal liability. Miss. Law Rev. 85, 7.