Inference from Biased Polls*

Andy Brownback† Nathaniel Burke‡ Tristan Gagnon-Bartsch§

March 7, 2023

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' explicit responses can predict SDR.

JEL classification: D91, D84, D72.
Keywords: Polling, Social Desirability, Inference, Signaling, Selection Bias.

*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, and West Virginia University. 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).

†University of Arkansas: ABrownback@walton.uark.edu.
‡West Virginia University: nathaniel.burke@mail.wvu.edu
§Harvard University: gagnonbartsch@fas.harvard.edu.

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 predict and prepare for changes in government policies despite the 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 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 inference 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

[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).

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 experiment 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 (either 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—revealing 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. 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 action.[3] 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 are virtuous or stigmatized. 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 Children's Hospital 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,

[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] In total, we consider eight different actions. Six 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. We also consider stealing $1 from another participant in the study and taking $1 for yourself from a planned donation to the Make-A-Wish Foundation.

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.[4] 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 ($p < 0.01$).[5]

In social interactions, people often have experience both reporting their views to others—similar to our H group—and drawing inference based on others' reports—similar to our predictors. For this reason, we designed our study to additionally 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.

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 whether the signal they receive reflects "perception-inflating" boasting—which should be discounted—or reflects "perception-deflating" confessing—which should be given additional weight. We define a signal to be perception inflating if it implies

[4] 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 a key question about how experience with SDR affects predictor behavior.

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

greater social desirability *than the predictor initially guessed*. A signal is perception deflating if it implies lesser social desirability *than the predictor initially guessed*.[6] A perception-deflating 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 perception-inflating signals from the H group by 18% relative to the IC group ($p < 0.001$), they treat perception-deflating 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.[7]

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

[6] These definitions facilitate a within-individual analysis: whether a given signal would increase or decrease a predictor's updated guess will depend on that predictor's initial guess, and hence the directional effect of a given signal will differ across individuals. There are no objectively high or low signals, only higher or lower signals than initial guesses. Under this definition, even a perception-deflating signal from the H group is consistent with SDR. SDR suggests that more people hypothetically claim socially desirable behavior than actually choose it, while a perception-deflating signal shows fewer people claiming socially desirable behavior relative to *the predictor's initial guess* of how many would actually choose it.

[7] A similar pattern emerges in experiments on "cursed thinking" (Eyster and Rabin, 2005) in trading environments with asymmetric information. Subjects often accept financial trades with better-informed parties to their own detriment (e.g., Samuelson and Bazerman, 1985). However, when explicitly asked, a typical subject in such settings correctly predicts that her better-informed partners will only agree to trades that are detrimental for her to accept (Hales, 2009).

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.

Given the limited social sophistication that we find, the benefits of collecting and disseminating more accurate data are clear. 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.[8] 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 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.[9]

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 (2021) 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 (2021) 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

[8] 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).

[9] "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.

between the two studies may be informative about the mechanisms at play. Braghieri (2021) 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 (2021) exhibit greater sophistication than subjects in our study.[10]

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 to misreporting their private information. Although experimental studies demonstrate 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.

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 stages: the Sentiment Stage, the Choice Stage, and the Prediction Stage. Each stage took place online with subjects recruited from the University of Arkansas. Each stage featured the same eight actions framed as binary choices.

We recruited 39 subjects for the Sentiment Stage. For the Choice Stage, we recruited 187 subjects and split them into two groups. In the Prediction Stage, we recruited 95 new subjects to combine with returners from the Choice Stage.

2.1 Actions

Subjects considered eight binary choices to take an action or not. The eight actions were:

St Jude Donation: Donate $1 to the St. Jude Children's Hospital.

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

NPR Donation: Donate $1 to KUAF radio station, the local NPR affiliate.

Steal: Steal $1 from a participant in another stage 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.

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

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

We made no attempt to label each action as "virtuous" and "stigmatized" based on our a priori perceptions. We designed our experiment and all hypotheses to be agnostic about the sentiment surrounding actions; instead, we classify actions based solely on the evaluations of the sentiment group, who are drawn from the same population. In this way, all of our tests could be based on perceptions that are observably present in the population.

Though the emotional valence of a specific action was unimportant to our design, it was important that the actions we selected possessed a wide range of emotional valence so that we could test for sensitivity to *differences* in social desirability. It was also important that the actions did not exist solely at the extremes of virtue and stigma so that we could identify effects both *between* and *within* stigmatized and virtuous domains. 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 towards stigmatized and virtuous actions.[11]

The binary nature of decisions (either 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 stages of the study, including information about the anonymity under which choices and statements were made. See Appendix Section C for the full description given to subjects.

[11] Moreover, we selected several actions related to political views since this is a familiar domain in which people observe poll data.

2.2 Sentiment Stage

We recruited 39 subjects to evaluate the sentiment associated with each of the eight actions listed above. Subjects who participated in the Sentiment Stage 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: $V_{i,A} \equiv \frac{\sum_{j=1}^{3} Q_{i,j,A}}{3}$.[12] Letting N_S denote the number of subjects in the Sentiment Stage, we will use the following indices to measure the perceived virtue of action A:

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

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

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 Stage

In the Choice Stage, 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

[12] 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 A.1 for details.

through claims in a hypothetical elicitation. The IC group had 91 subjects and the H group had 96 subjects.[13]

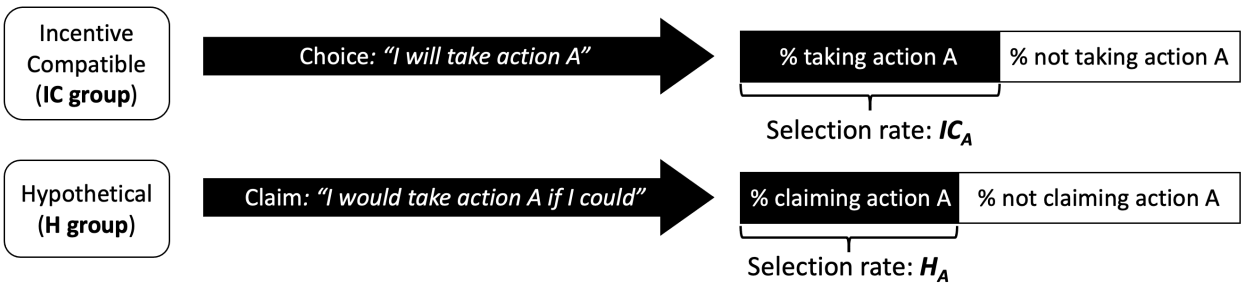
The only difference between the IC and H groups was the incentive-compatibility of the IC group's choice elicitation. For instance, if a subject in the IC group chose to donate \$1 to St. Jude, then that subject actually sacrificed \$1 of their payment and St. Jude actually received a \$1 donation. If 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 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. Importantly, SDR_A can take on negative values, indicating a socially undesirable action. Figure 1 depicts the flow of the Choice Stage for an example where the H group understates demand for an action (i.e. $SDR_A < 0$).

Figure 1. Experimental Design: Choice Stage



All subjects received a \$5 participation payment in the Choice Stage. 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

[13] We restricted subjects to participate in the Choice Stage only once—either in the IC or H group. We dropped 15 submissions from the IC group and 7 from the H group for violating this restriction. 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.

the Choice Stage were told that they must participate in an additional stage (the Prediction Stage, described below) during which they could earn more money; subjects were not given any description of this additional stage until the Prediction Stage began.

2.4 Prediction Stage

In order to receive their full payment, all of the subjects who participated in the Choice Stage were required to participate as "predictors" in the Prediction Stage, which started five days later. In addition, we recruited 95 new predictors who had not participated in any previous stage. In total, the Prediction Stage 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 stage along with any earnings gained from accurate predictions.

In the Prediction Stage, predictors observed the exact same descriptions of the actions as subjects in the Choice Stage and were tasked with guessing the choice behavior of the IC group for each of the actions.[14] 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).[15]

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 $\textsc{Guess}_{i,1,A}$ 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.[16] 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

[14] As mentioned in Subsection 2.3 (Footnote 13), some subjects violated the restriction for duplicate participation. We discovered these duplicates after the Prediction Stage, 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.

[15] 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).

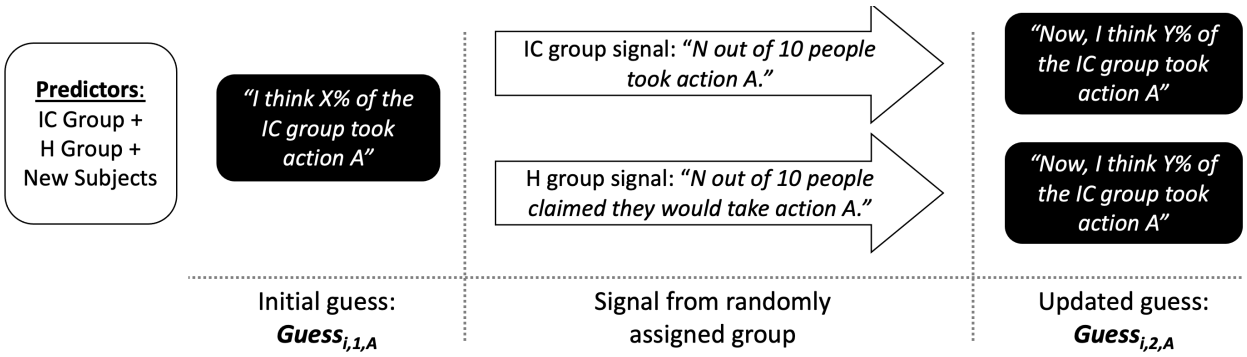
[16] More specifically, $s_{i,A}$ counts the number of 10 randomly-chosen respondents who elected to take action A. A predictor received such a signal for each action, and thus received 8 signals in total. A predictor received all signals from the same group: either 8 signals from the IC group or 8 from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

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.[17]

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

Figure 2. Experimental Design: Prediction Stage



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.

2.5 Recruitment Summary

Table 1 breaks our sample down by assignment.

Table 1. Subject participation by treatment

	Sentiment Stage	Choice Stage	Prediction Stage
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 Stage after successfully completing the Choice Stage.

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

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.

Recall that the H group faces no incentives based on their claims. Hence, it is costless for these subjects to claim that they would take the 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.

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 socially desirable way regardless of their true preference.[18] If action A is 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 reveal 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 $SDR_A = H_A - IC_A = \theta_A(1-IC_A)$.[19]

If action A is instead stigmatized, then H-group subjects will deflate their claimed desire to take the action. Their expected selection rate is thus $H_A = (1-\theta_A)IC_A$: a fraction $1-\theta_A$ of subjects again reveal 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

[18] 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).

[19] This formulation assumes that if a person's hypothetical behavior deviates from their incentivized behavior, then it must deviate in the direction of social desirability. However, deviations may also occur due to noise. This noise can easily be incorporated into our approach, and we discuss the interpretation of our results in light of such potential noise in Section 5.

10 random subjects from either the IC or H group. If predictor i is assigned to receive information from the IC group, then $s_{i,A} \sim \text{Bin}(10, IC_A)$. If predictor i instead receives information from the H group, then our model implies that, if action A is virtuous, then $s_{i,A} \sim \text{Bin}(10, (1-\theta_A)IC_A + \theta_A)$, and if it is stigmatized, then $s_{i,A} \sim \text{Bin}(10, (1-\theta_A)IC_A)$. Although the distribution of signals depends on θ_A, we do not assume perfect knowledge of θ_A in developing our hypotheses about social sophistication. 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.[20]

We now present the hypotheses that we test in Sections 4 and 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 increasingly overstate claimed demand for an action as 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. In the model above, if an action A' is perceived to be more virtuous than action A, then $\theta_{A'} > \theta_A$. That is, more subjects are inclined to lie in the socially desirable way when incentives are removed. The measures of SDR for both virtuous and stigmatized actions, derived above, are increasing in θ_A.[21]

[20] Below, we will impose an additional key assumption: although signals may influence a predictor's estimate of θ_A, they must also influence their beliefs about IC_A.

[21] This can be thought of as a local phenomenon, as it assumes that choice rates, IC_A, are held constant as perceived virtue changes. This is one limitation of such a stylized model, because a sufficiently large change in the virtue or stigma of an action would likely affect choice rates. However, this should have little impact on our results, as our actions are all within a reasonable range of stigma or virtue. Following the logic of our model provides useful intuition in this context.

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.

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 a sophisticated understanding of θ_A *should* influence the way predictors respond to 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.[22] 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.

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

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

[22] 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.

sophistication, updated guesses about the behavior of the IC group will react more strongly to signals from the IC group.

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.

Hypothesis 2 (Direction): *Predictors with social sophistication will discount "perception-inflating" signals from the H group relative to similar signals from the IC group, but they will give greater relative weight to "perception-deflating" signals from the H group.*[23]

The effect of θ_A on predictions depends on whether action A is stigmatized or virtuous. Thus, sophisticated inference requires a predictor to first assess whether an action is virtuous (where θ_A correlates with overstatement of claimed demand by the H group) or stigmatized (where θ_A correlates with understatement of claimed demand by the H group). Predictors can then categorize the signal they observe as a perception-inflating boast or a perception-deflating confession. Since there are no objectively "high" or "low" signals, we define perception-inflating and perception-deflating signals relative to a predictor's initial guesses. A predictor's signal is perception-inflating (-deflating) if it indicates greater (lesser) demand for socially-desirable actions *than the predictor initially guessed.* Full social sophistication requires a heterogeneous treatment of perception-inflating and perception-deflating signals. Sophisticated predictors should discount perception-inflating signals from the H group as they are likely over-optimistic about socially-desirable choices. But, in the rare event that a predictor observes a perception-*deflating* signal from the H group, this signal should be given *more* weight than an equivalent signal from the IC group.[24] This is because a sophisticated predictor realizes that they have observed a perception-deflating signal *despite* the H group's ability to costlessly overstate their claimed demand for socially desirable behavior. Thus, IC-group choices are probably even lower than this signal suggests.

For example, donations to St. Jude are categorized as virtuous because $SDR_{St.Jude} > 0$. Now, 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 perception-inflating. 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

[23] This hypothesis was not included in our pre-analysis plan. We include it here and provide results in the following section 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.

[24] SDR predicts that perception-inflating signals will be more likely from the H group than the IC group. Indeed, perception-inflating signals are 16 percentage points more likely when signals arrive 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*.

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 perception-deflating. 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]

To summarize this test in terms of our stylized model, we jointly test if predictors (i) identify whether θ_A inflates or deflates hypothetical claims about a given action A and (ii) understand that this makes perception-deflating signals from the H group less likely and, therefore, more informative about IC_A.

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

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, the boundary case of $\theta_A = 1$: signals from the H group then provide no information and should be ignored. Social sophistication suggests that a predictor should account for the relative magnitude of θ_A across actions (i.e. which actions have greater or lesser degrees of virtue or stigma). Thus, as the perceived virtue or stigma of an action grows relatively more extreme, sophisticated predictors should increase their discounting of H-group signals relative to IC-group signals.

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

4 Data Description and Manipulation Checks

In this section, we provide summary statistics for each action in each stage 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

[25] This assumes that the predictor does not use their signal to infer whether the action is stigmatized or virtuous. If this were the case, then a sophisticated predictor may use the surprisingly low 40% signal from the H group to conclude that donations to St. Jude are in fact *stigmatized*. Results from the Sentiment Stage support our assumption, showing that actions have predictable stigma or virtue.

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 Stage and Choice Stage are captured in Columns 1 and Columns 2–3, respectively. Initial and updated guesses from the Prediction Stage 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.

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.

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 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 virtue, which we measured independently during the Sentiment Stage.

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 Stage (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 virtue and stigma from the Sentiment Stage. 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

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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

deviation increase in perceived virtue.[26]

The fact that our subjects' evaluations of sentiment are predictive of observed SDR demonstrates their "sentiment sophistication:" they have a fairly accurate understanding of the stigma or virtue surrounding an action. Using the knowledge of which actions are more socially desirable—and therefore more likely to inspire dishonest responding 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 are able to 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.

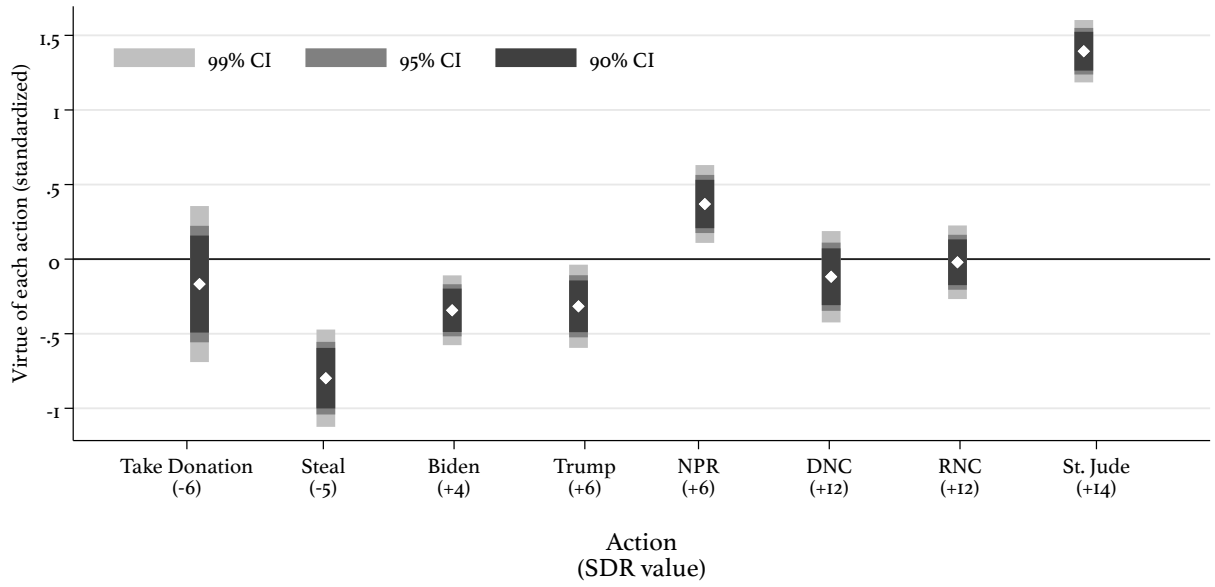
This sentiment sophistication is presented graphically in Figure 3, which orders each of the eight actions along the horizontal axis according to their observed SDR. For each action, the associated sentiment evaluations are plotted on the vertical axis, revealing a clear positive association between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

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

[26] 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.

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



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: $(IC_A - \text{GUESS}_{i,t,A})^2$. Since predictors were randomly assigned to receive signals from either the IC or H group after stating their initial guess, the baseline accuracy was balanced.[27] 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.

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.

[27] The p-values for a test of differences in the accuracy of initial guesses are $p = 0.97$ and $p = 0.81$ for absolute- and squared-errors, respectively.

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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

82% of subjects provide updated guesses that exhibit this "between-ness" property for at least six of the eight actions, while 42% of them exhibit it for all eight actions.[28]

Although updated guesses do improve in accuracy, predictors in both groups still fall short of a simple heuristic: the accuracy of their signals. Both groups would improve their accuracy by simply making guesses that match their signals exactly.[29] 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 test each hypothesis, we will begin with our pre-registered specification before presenting any alternative specifications. Appendix Section B details

[28] As discussed in Section 3, deriving Bayesian benchmarks is not possible without eliciting subjects' entire prior belief distribution over IC choice rates.

[29] 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 $(10 \times z)\%$.

our empirical estimation and the ways in which we supplement our pre-registered analysis.

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 Stage, and thus the distribution of signals differs across groups. Therefore, our random assignment of information source is confounded with the assignment of a different mean for the distribution of signals. To resolve this confound and isolate the random sampling variation in signals, we control for the differences in the distributions from which signals are drawn.[30] 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 explore the role of experience in prompting skepticism toward claims from the H group.

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

[30] See Kahan (2015) and Thaler (2019) 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
Signal Value	0.59***	0.55***	0.55***	0.54***
	(0.03)	(0.05)	(0.07)	(0.06)
IC Info Source × Signal Value	0.08***	0.17***	0.11	0.22**
	(0.03)	(0.07)	(0.10)	(0.09)
Initial Guess	0.30***			
	(0.02)			
IC Info Source	-1.14			
	(1.07)			
Observations	2168	2168	760	1408
Clusters	271	271	95	176
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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) 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 Stage 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 Stage is a source of skepticism toward H-group signals. Predictors who previously participated in the Choice Stage 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 Stage approaches marginal significance ($p = 0.125$). These results can be found in Appendix Section A.3.[31]

These findings could be driven in part by predictors believing that H signals were less reliable due to inattentive or noisy hypothetical responses. Our manipulation checks 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, leading them to discount these 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 signals to be substantially noisier than IC signals, they should have less confidence in their updated guesses after receiving H signals. As we will see below, predictors do not demonstrate these patterns of confidence in their updated guesses.

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 predictors' guesses appreciate the direction in which SDR will affect signals from the H group. That is, we ask how accurately predictors recognize whether the H group will tend to overstate or understate their claimed desire to take a given action.

Predictors with social sophistication should discount signals from the H group more when they are "perception-inflating"—i.e., suggestive of more socially desirable behavior than the predictor's initial guess—because the H-group signals are drawn from claims that tend to be optimistic exaggerations. Conversely, social sophistication guides predictors to give *more* weight to signals from the H group when they are "perception-deflating"—i.e., suggestive of less socially desirable behavior—because when a pessimistic signal is drawn from the H

[31] Table A.3 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.

group's optimistic claims, the actual choices of the IC group are likely even more pessimistic. For a concrete example and further details on this logic, see the discussion of Hypothesis 2 in Section 3.

To evaluate this hypothesis, we must first designate which actions are socially desirable. We do so empirically using SDR_A. If $SDR_A > 0$—that is, the H group overstates their demand for action A—then A is considered virtuous. Otherwise, if $SDR_A < 0$, then A is considered stigmatized.[32,33] With knowledge of an action's social desirability, social sophistication will enable predictors to determine if the signal they receive is perception-inflating or perception-deflating. A perception-inflating signal is one that indicates a greater demand for an action with $SDR_A > 0$ (or a lesser demand for an action with $SDR_A < 0$) *than the predictor initially guessed*. A perception-deflating signal indicates a lesser demand for an action with $SDR_A > 0$ (or a greater demand for an action with $SDR_A < 0$) *than the predictor initially guessed.*

Note that absolute thresholds for perception-inflating (-deflating) signals do not exist; we can only categorize them based on whether they indicate greater (lesser) social-desirability than the predictor's initial guess. As long as initial guesses are uncorrelated with predictor traits, then our between-group analysis is as cleanly-identified as our within-group analysis.

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 perception-inflating or perception-deflating. To aid the interpretation of coefficients, we will replicate the analysis of Hypothesis 1 separately for predictors receiving perception-inflating and perception-deflating signals.[34]

Table 6 displays our limited support for Hypothesis 2. Columns 1 and 2 replicate the analysis of Columns 1 and 2 from Table 5 but restrict their focus to perception-inflating 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 0.11 percentage points less when the signals arrive from the H group ($p < 0.001$). Column 2 estimates this diminished weight to be 0.12 ($p = 0.211$). Thus, predictors do appear to recognize that perception-inflating 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.

[32]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.

[33]One could consider using our measure of perceived virtue from the Sentiment Stage to identify stigmatized actions. However, this approach presents an issue with units because the Likert scale we used to measure sentiment does not have an obvious cutoff for socially-desirable and socially-undesirable actions.

[34]We conduct similar analysis using a fully-interacted specification in Appendix Section A.4. The results are qualitatively similar, but even more inconsistent with social sophistication.

Table 6. Updated guesses in response to perception-inflating and -deflating signals from different sources

	Perception-Inflating		Perception-Deflating	
	Updated Guess	Δ Guess	Updated Guess	Δ Guess
Signal Value	0.61***	0.26***	0.71***	0.33***
	(0.05)	(0.07)	(0.03)	(0.06)
IC Info Source×Signal Value	0.11***	0.12	-0.01	0.08
	(0.04)	(0.09)	(0.03)	(0.09)
IC Info Source	-2.06		0.67	
	(1.27)		(1.28)	
Initial Guess	0.29***		0.23***	
	(0.04)		(0.03)	
Observations	925		1243	
Clusters	267		270	
Control for Mean Signal:	Yes	N/A	Yes	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. "Perception-Inflating" ("Perception-Deflating") are defined by whether the signal is in the direction of more (less) social desirability relative to the initial guess. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Social sophistication also requires recognizing that perception-deflating signals should be given relatively *more* weight when they come from the H group. Columns 3 and 4 show no evidence for this more complex dimension of social sophistication. In Column 3, instead of the predicted negative coefficient for "IC Info Source×Signal Value," we find it to be near zero and insignificant, revealing no difference in the weight given to H-group signals. Additionally, in the specification of Column 4, this effect is positive, meaning that predictors still discount signals from the H group even when they are perception-deflating, though this effect is not significant. Thus, we conclude that predictors' guesses demonstrate no recognition that perception-deflating signals from the H group are even stronger indictments of behavior than corresponding signals from the IC group.[35]

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 Figure 3 show how Sentiment Stage 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.

[35] In Appendix Section A.4, we present individual guesses in Figures A.3 and A.4 to visualize heterogeneous discounting with respect to the direction of SDR. These figures mirror the approach taken in Figures A.1 and A.2 (which visualize average discounting).

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 Stage as the interaction term. Table 3 and Figure 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 and 2, 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.

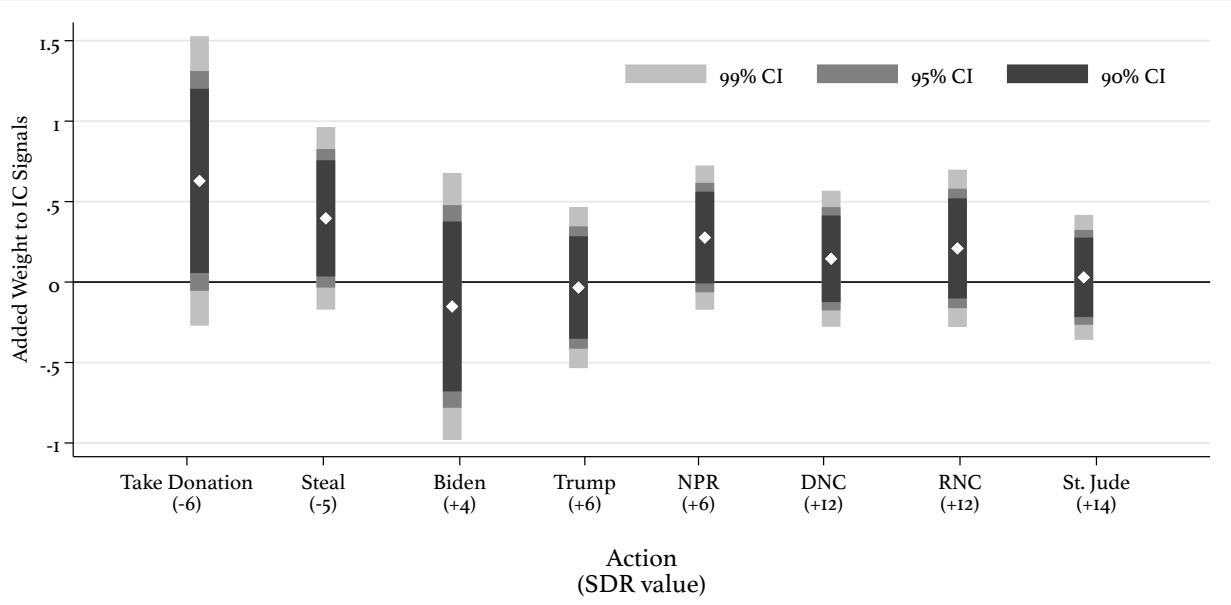
Table 7. Updated guesses in response to SDR magnitude by information source

	Updated Guess	Δ Guess	Δ Guess
Signal Value	0.60***	0.57***	0.63***
	(0.05)	(0.12)	(0.06)
IC Info Source×Signal Value	0.06	0.38**	0.24**
	(0.07)	(0.17)	(0.09)
Signal Value×$\lvert SDR\rvert$	-0.00	-0.00	
	(0.00)	(0.01)	
IC Info Source×Signal Value×$\lvert SDR\rvert$	0.00	-0.02	
	(0.01)	(0.02)	
IC Info Source×$\lvert SDR\rvert$	-0.14	-0.46	
	(0.16)	(0.47)	
$\lvert SDR\rvert$	0.17	-0.40	
	(0.13)	(0.90)	
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	
Clusters		271	
Control for Mean Signal:	Yes	N/A	N/A
Fixed-Effects:	Action	Action×Source	Action×Source

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

Figure 4 visualizes this lack of social sophistication.[36] As in Figure 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 Figure 3—which demonstrated clear sentiment sophistication—Figure 4 finds no evidence of social sophistication.

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



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 Figure 4, predictors fail this test. This failure is striking because the test is a natural extension of the tests from Table 3 and Figure 3. In those, the sentiment group demonstrates a clear understanding of which actions tend to be more virtuous or stigmatized. 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.

[36] Figure 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.

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 on 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.[37] 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 standard deviation of 1.

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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Column 1 shows a false confidence by predictors. 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

[37] The analysis of predictor confidence is not included in our pre-analysis plan. However, our framework of analysis mirrors that of the pre-registered hypotheses, and we believe it substantively adds to our understanding of the impact of (perceived) information quality.

and confidence is diminished and statistically insignificant for predictors who receive higher-quality information from the IC group ($p > 0.40$).[38] 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, they 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 are discounting claims of the H group because they believe that information to be noisier.

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.

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 a clear ability to identify the social desirability of actions, predictors failed to translate this knowledge into sophisticated discounting. While our subjects correctly discounted perception-inflating signals, they incorrectly responded to perception-deflating signals. This is a failure of sophistication: subjects should appreciate that perception-deflating signals are especially informative because 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

[38] Gneezy and Serra-Garcia (2019) find similar overconfidence in one's ability to detect lies by others.

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.[39]

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.[40] 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 themselves. In this way, biased poll data may carry over into biased inferences and sub-optimal actions.

[39] 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.

[40] 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., 2021) 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. (2020) 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.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–1153.

Adkins, R. E. and A. J. Dowdle (2002): "The Money Primary: What Influences the Outcome of Pre-Primary Presidential Nomination Fundraising?" *Presidential Studies Quarterly*, 32, 256–275.

Agranov, M., J. K. Goeree, J. Romero, and L. Yariv (2018): "What makes voters turn out: The effects of polls and beliefs," *Journal of the European Economic Association*, 16, 825–856.

Arnold, H. J., D. C. Feldman, and M. Purbhoo (1985): "The role of social-desirability response bias in turnover research," *Academy of Management Journal*, 28, 955–966.

Augenblick, N. and M. Rabin (2019): "An experiment on time preference and misprediction in unpleasant tasks," *Review of Economic Studies*, 86, 941–975.

Bagwell, L. S. and B. D. Bernheim (1996): "Veblen effects in a theory of conspicuous consumption," *The American Economic Review*, 349–373.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964): "Measuring utility by a single-response sequential method," *Behavioral Science*, 9, 226–232.

Bénabou, R. and J. Tirole (2002): "Self-confidence and personal motivation," *The Quarterly Journal of Economics*, 117, 871–915.

Benjamin, D. J. (2019): "Errors in probabilistic reasoning and judgment biases," in *Handbook of Behavioral Economics: Applications and Foundations*, Elsevier, vol. 2, 69–186.

Bharadwaj, P., M. M. Pai, and A. Suziedelyte (2017): "Mental health stigma," *Economics Letters*, 159, 57–60.

Boukouras, A., W. Jennings, L. Li, and Z. Maniadis (2020): "Can Biased Polls Distort Electoral Results? Evidence From The Lab," Working paper, School of Business, University of Leicester.

Braghieri, L. (2021): "Political Correctness, Social Image, and Information Transmission," Working paper.

Brown-Iannuzzi, J. L., M. B. Najle, and W. M. Gervais (2019): "The illusion of political tolerance: Social desirability and self-reported voting preferences," *Social Psychological and Personality Science*, 10, 364–373.

Brownback, A. and A. Novotny (2018): "Social desirability bias and polling errors in the 2016 presidential election," *Journal of Behavioral and Experimental Economics*, 74, 38–56.

Bursztyn, L., D. Cantoni, P. Funk, and N. Yuchtman (2021): "Do Polls Affect Elections? Evidence from Swiss Referenda," Working paper.

Bursztyn, L., B. Ferman, S. Fiorin, M. Kanz, and G. Rao (2018): "Status Goods: Experimental Evidence from Platinum Credit Cards," *The Quarterly Journal of Economics*, 133, 1561–1595.

Cantú, F. and J. Márquez (2021): "The effects of election polls in Mexico's 2018 presidential campaign," *Electoral Studies*, 73, 102379.

Carlson, E. (2018): "The perils of pre-election polling: Election cycles and the exacerbation of measurement error in illiberal regimes," *Research & Politics*, 5, 2053168018774728.

Charness, G., R. Oprea, and S. Yuksel (2021): "How do people choose between biased information sources? Evidence from a laboratory experiment," *Journal of the European Economic Association*, 19, 1656–1691.

Coffman, K. B., L. C. Coffman, and K. M. M. Ericson (2017): "The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated," *Management Science*, 63, 3168–3186.

Crawford, V. (1998): "A survey of experiments on communication via cheap talk," *Journal of Economic theory*, 78, 286–298.

Crawford, V. P. and J. Sobel (1982): "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1431–1451.

de Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and bounding experimenter demand," *American Economic Review*, 108, 3266–3302.

Del Boca, F. K. and J. A. Noll (2000): "Truth or consequences: the validity of self-report data in health services research on addictions," *Addiction*, 95, 347–360.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual risk attitudes: Measurement, determinants, and behavioral consequences," *Journal of the european economic association*, 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," *Journal of the European Economic Association*, 9, 43–60.

Eyster, E. and M. Rabin (2005): "Cursed equilibrium," *Econometrica*, 73, 1623–1672.

Farrell, J. and M. Rabin (1996): "Cheap talk," *Journal of Economic perspectives*, 10, 103–118.

Fedyk, A. (2018): "Asymmetric naivete: Beliefs about self-control," *Available at SSRN 2727499*.

Finkel, S. E., T. M. Guterbock, and M. J. Borg (1991): "Race-of-interviewer effects in a preelection poll Virginia 1989," *Public Opinion Quarterly*, 55, 313–330.

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in Disguise—An Experimental Study on Cheating," *Journal of the European Economic Association*, 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.

Gneezy, U. and M. Serra-Garcia (2019): "Mistakes and Overconfidence in Detecting Lies," Working paper.

Gonzalez-Ocantos, E., C. K. De Jonge, C. Meléndez, J. Osorio, and D. W. Nickerson (2012): "Vote buying and social desirability bias: Experimental evidence from Nicaragua," *American Journal of Political Science*, 56, 202–217.

Grosser, J. and A. Schram (2010): "Public opinion polls, voter turnout, and welfare: An experimental study," *American Journal of Political Science*, 54, 700–717.

Haaland, I., C. Roth, and J. Wohlfart (2020): "Designing information provision experiments," Working Paper 20/20, CEBI Working Paper Series.

Hales, J. (2009): "Are investors really willing to agree to disagree? An experimental investigation of how disagreement and attention to disagreement affect trading behavior," *Organizational Behavior and Human Decision Processes*, 108, 230–241.

Heerwig, J. A. and B. J. McCabe (2009): "Education and social desirability bias: The case of a Black presidential candidate," *Social Science Quarterly*, 90, 674–686.

Heidhues, P. and B. Kőszegi (2010): "Exploiting naivete about self-control in the credit market," *American Economic Review*, 100, 2279–2303.

Holbrook, A. L. and J. A. Krosnick (2010): "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique," *Public Opinion Quarterly*, 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," *The Journal of Politics*, 71, 769–781.

Janus, A. L. (2010): "The Influence of Social Desirability Pressures on Expressed Immigration Attitudes," *Social Science Quarterly*, 91, 928–946.

Jones, A. E. and M. Elliot (2016): "Examining Social Desirability in Measures of Religion and Spirituality Using the Bogus Pipeline," *Review of Religious Research*, 1–18.

Kahan, D. M. (2015): "The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it," *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, 1–16.

Kane, J. G., S. C. Craig, and K. D. Wald (2004): "Religion and presidential politics in Florida: A list experiment," *Social Science Quarterly*, 85, 281–293.

Karlan, D. S. and J. Zinman (2012): "List randomization for sensitive behavior: An application for measuring use of loan proceeds," *Journal of Development Economics*, 98, 71–75.

Kartik, N. (2009): "Strategic communication with lying costs," *The Review of Economic Studies*, 76, 1359–1395.

Kruger, J. and D. Dunning (1999): "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of Personality and Social Psychology*, 77, 1121.

Krupka, E. L. and R. A. Weber (2013): "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11, 495–524.

Krysan, M. (1998): "Privacy and the expression of white racial attitudes: A comparison across three contexts," *Public Opinion Quarterly*, 506–544.

Latkin, C. A., C. Edwards, M. A. Davey-Rothwell, and K. E. Tobin (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," *Addictive Behaviors*, 73, 133–136.

Maccoby, E. E. and N. Maccoby (1954): "The interview: A tool of social science," *Handbook of Social Psychology*, 1, 449–487.

Paulhus, D. L. (1984): "Two-component models of socially desirable responding," *Journal of Personality and Social Psychology*, 46, 598–609.

Powell, R. J. (2013): "Social desirability bias in polling on same-sex marriage ballot measures," *American Politics Research*, 41, 1052–1070.

Pronin, E., D. Y. Lin, and L. Ross (2002): "The bias blind spot: Perceptions of bias in self versus others," *Personality and Social Psychology Bulletin*, 28, 369–381.

Raghavarao, D. and W. T. Federer (1979): "Block total response as an alterNative to the randomized response method in surveys," *Journal of the Royal Statistical Society. Series B (Methodological)*, 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., K. Imai, and J. N. Shapiro (2016): "An empirical validation study of popular survey methodologies for sensitive questions," *American Journal of Political Science*, 60, 783–802.

Samuelson, W. and M. Bazerman (1985): "The Winner's Curse in Bilateral Negotiations," in *Research in Experimental Economics*, ed. by V. Smith, Greenwich, CT: JAI Press, vol. 3, 105–137.

Stephens-Davidowitz, S. (2014): "The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data," *Journal of Public Economics*, 118, 26–40.

Streb, M. J., B. Burrell, B. Frederick, and M. A. Genovese (2008): "Social Desirability Effects and Support for a Female American President," *Public Opinion Quarterly*, 72, 76–89.

Thaler, M. (2019): "The "Fake News" Effect: An Experiment on Motivated Reasoning and Trust in News," Working paper.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000): *The psychology of survey response*, Cambridge University Press.

Tourangeau, R. and T. Yan (2007): "Sensitive questions in surveys." *Psychological bulletin*, 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," *Journal of the American Statistical Association*, 60, 63–69.

West, R. F., R. J. Meserve, and K. E. Stanovich (2012): "Cognitive sophistication does not attenuate the bias blind spot." *Journal of personality and social psychology*, 103, 506–519.

Yeargain, T. (2020): "Fake Polls, Real Consequences: The Rise of Fake Polls and the Case for Criminal Liability," *Missouri Law Review*, 85, 7.

A Appendix A: Supplemental Analysis

A.1 Breakdown 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 indicates 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 Stage. While these relationships are all positive and most are significant, there appears to be a stronger association between anticipation of others' sentiment (Columns 5–6) rather than own-sentiment (Columns 1–2) or sentiment towards 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.

Table A.1. Socially desirable responding and perceived virtue

	Socially Desirable Responding					
	Measure 1		Measure 2		Measure 3	
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.2 Heterogeneous Responses to Signals

The analysis in Table 5 is limited to aggregate updating and could obscure important heterogeneity in updating behavior. Figures 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 Figure A.1.[41] 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.2 tests whether these behaviors—in addition to partial updating—differ across information sources.

Table A.2 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 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 ($p = 0.023$).

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

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

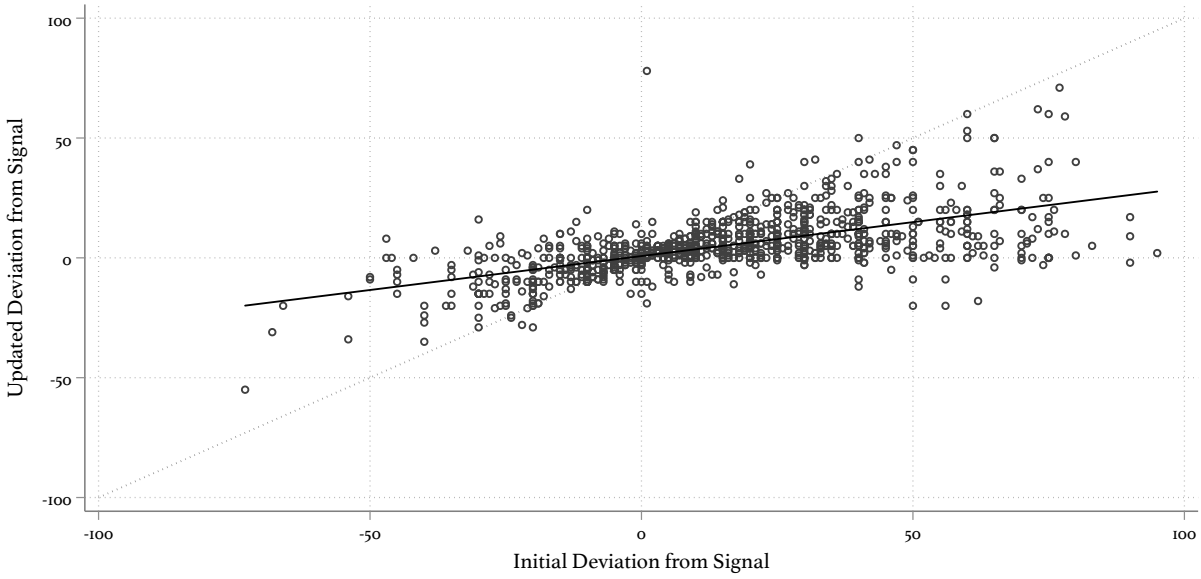


Figure A.2. Predictors receiving signals from the H group

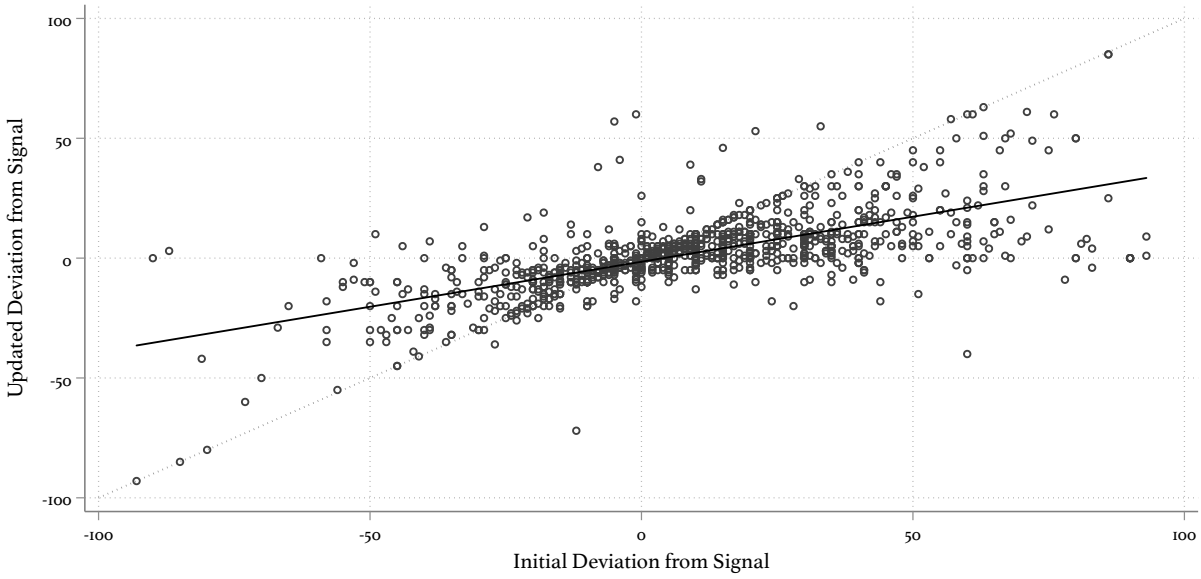


Table A.2. Updated guesses in response to signals from different sources

	Pr[Ignore Signal]	Pr[Match Signal]	Δ Guess (partial updating)
IC Info Source	-0.03	0.03	
	(0.02)	(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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3 Experience with SDR

To examine the mechanisms driving social sophistication, we explore whether predictors who previously participated in the Choice Stage are better at accounting for SDR in poll data than newly recruited predictors. A predictor with experience in the Choice Stage 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 Stage, (ii) those who participated in the H group in the Choice Stage, and (iii) newly recruited predictors who did not participate in the Choice Stage. 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 Stage 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.

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 $p = 0.85$).

Table A.3. Updated guesses by information source across groups with different prior experience

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

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

A.4 Direction of SDR

In Table A.4, we replicate the analysis of Section 5.2 using a fully-interacted model. This approach produces qualitatively similar results as seen in the coefficients for "Perception-Inflating×IC Info Source×Signal Value" and "Perception-Deflating×IC Info Source×Signal Value." We again find that predictors discount perception-inflating signals from the H group. However, the incorrect discounting of perception-deflating signals from the H group are now significant in the within-subjects specification. Thus, behavior is even less consistent with social sophistication under this approach.

Figures A.3 and A.4 replicate the visualizations from Figures 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 pre-

Table A.4. Updated guesses in response to perception-inflating signals from different sources

	Updated Guess	Δ Guess
Perception-Inflating×Signal Value	0.55***	0.55***
	(0.03)	(0.05)
Perception-Inflating×IC Info Source×Signal Value	0.16***	0.29***
	(0.04)	(0.08)
Perception-Inflating×IC Info Source	-3.14**	-8.60**
	(1.41)	(3.46)
Perception-Inflating	1.01	1.71
	(1.11)	(2.09)
Perception-Deflating×Signal Value	0.68***	0.41***
	(0.03)	(0.05)
Perception-Deflating×IC Info Source×Signal Value	-0.01	0.15**
	(0.03)	(0.08)
Perception-Deflating×IC Info Source	0.85	-1.51
	(1.15)	(2.42)
Initial Guess	0.30***	
	(0.02)	
Observations	2168	
Clusters	271	
Control for Mean Signal:	Yes	N/A
Fixed-Effects:	Action	Action×Source

Notes: Random-effects linear regression with subject-level random effects. Standard errors clustered at the individual level. "Perception-Inflating" ("Perception-Deflating") are indicator variables equal to one if the signal is in the direction of more (less) social desirability relative to the initial guess. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

dictions about the number of subjects who steal from another subject with the number of subjects who refuse to steal from another subject.

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

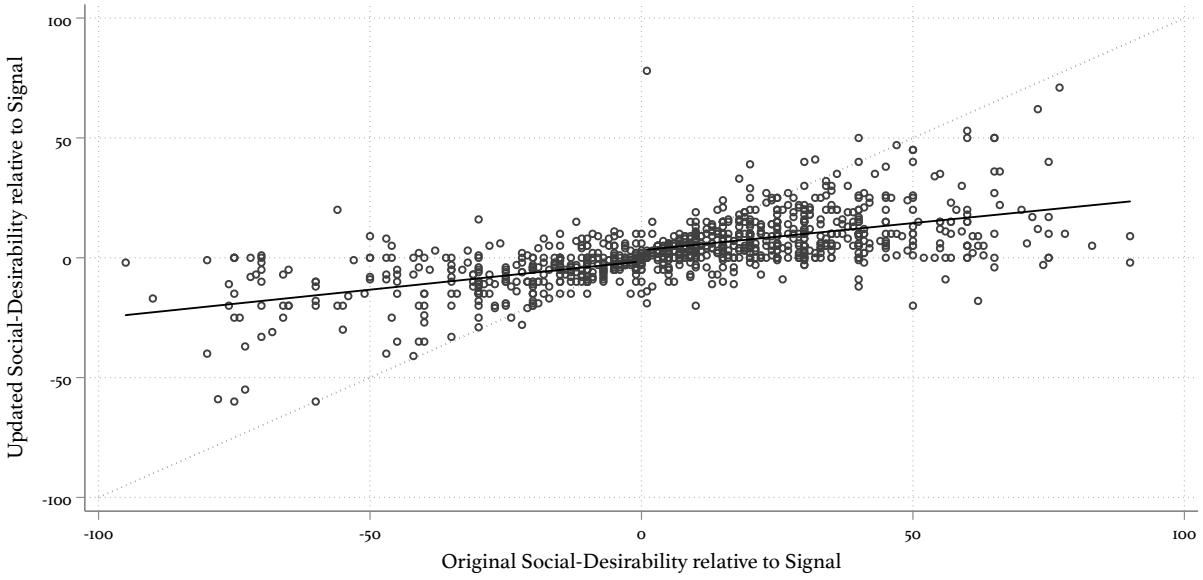
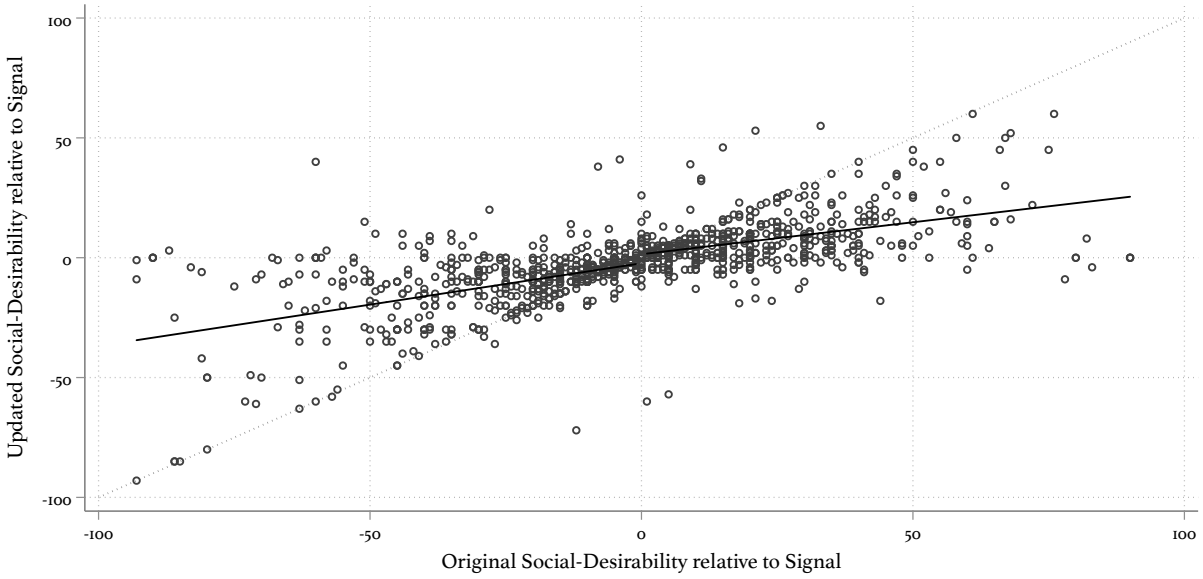


Figure A.4. Predictors receiving signals from the H group



Figures 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 Figure A.3 than on the left side of Figure A.4.

B Appendix B: Details on Analysis

This appendix provides details about the specific regressions underlying our results in Sections 4–**??**. 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–**??**, 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:

$$SDR_A = \beta_0 + \beta_1 \times V_A + \epsilon_A, \tag{4}$$

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

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:

$$SDR_A = \beta_0 + \beta_1 \times \widehat{V}_{i,A} + \nu_i + \epsilon_{i,A}. \tag{5}$$

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:

$$ABS_{i,2,A} = \beta_0 + \beta_1 ABS_{i,1,A} + \beta_2 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \tag{6}$$

$$SQ_{i,2,A} = \beta_0 + \beta_1 SQ_{i,1,A} + \beta_2 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \tag{7}$$

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:

$$\Delta ABS_{i,A} = \beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}, \tag{8}$$

$$\Delta SQ_{i,A} = \beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}. \tag{9}$$

B.3 Hypothesis 1: Anticipation of SDR

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

$$\textsc{Guess}_{i,2,A} = \beta_0 + \beta_1 \textsc{Guess}_{i,1,A} + \beta_2 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \beta_5 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}, \tag{10}$$

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: $\Delta\textsc{Guess}_{i,A} = \textsc{Guess}_{i,2,A} - \textsc{Guess}_{i,1,A}$. 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:

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

Alternative Specification: Extensive- and Intensive-Margin Responses

Table A.2 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:

$$\Pr\left(\textsc{Match}_{i,A}\right) = \Phi\left(\beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}\right), \tag{12}$$

$$\Pr\left(\textsc{Ignore}_{i,A}\right) = \Phi\left(\beta_0 + \beta_1 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A}\right), \tag{13}$$

$$\Delta\textsc{Guess}_{i,A} = \beta_0 + \beta_1 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \delta_{T,A} + \nu_i + \epsilon_{i,A}, \tag{14}$$

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

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 perception-inflating—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, a signal is perception-inflating if it suggests that there are more people engaging in (or claiming to engage in) this action than the predictor initially guessed. The opposite is true for actions that are socially undesirable (i.e. $SDR_A < 0$).

Column 1 of Table 6 presents our first test of Hypothesis 2 using the same random-effects linear-regression specification as in Equation 10, but including a full set of interactions with terms that indicate whether the signal is perception-inflating or perception-deflating:

$$\begin{aligned}\textsc{Guess}_{i,2,A} = \beta_0 &+ \beta_1 \textsc{Guess}_{i,1,A} + \beta_2 S_{i,A} \times \text{PI}_{i,A} \\ &+ \beta_3 S_{i,A} \times \text{IC}_i \times \text{PI}_{i,A} + \beta_4 \text{IC}_i \times \text{PI}_{i,A} + \beta_5 \text{PI}_{i,A} + \beta_6 S_{i,A} \times \text{PD}_{i,A} \\ &+ \beta_7 S_{i,A} \times \text{IC}_i \times \text{PD}_{i,A} + \beta_8 \text{IC}_i \times \text{PD}_{i,A} + \beta_9 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}. \end{aligned} \tag{15}$$

Here, we interact all of the relevant terms from Equation 10 with $\text{PI}_{i,A}$ ($\text{PD}_{i,A}$), indicators for whether the signal is perception-inflating (perception-deflating) relative to $\textsc{Guess}_{i,1,A}$. 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 11, 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 10. However, we now include terms interacted with the absolute value of our measure of SDR:

$$\begin{aligned}\text{GUESS}_{i,2,A} &= \beta_0 + \beta_1 \text{GUESS}_{i,1,A} + \beta_2 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \beta_5 S_{i,A} \times |\text{SDR}_A| \\ &+ \beta_6 S_{i,A} \times \text{IC}_i \times |\text{SDR}_A| + \beta_7 \text{IC}_i \times |\text{SDR}_A| + \beta_8 |\text{SDR}_A| + \beta_9 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A}. \quad (16)\end{aligned}$$

Here, we interact all of the relevant terms from Equation 10 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 11, 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:

$$\widehat{\text{Confidence}}_{i,1,A} = \beta_0 + \beta_1 ABS_{i,1,A} + \delta_A + \nu_i + \epsilon_{i,A}, \tag{17}$$

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:

$$\begin{aligned}\widehat{\text{Confidence}}_{i,2,A} = \beta_0 + \beta_1 ABS_{i,1,A} + \beta_2 ABS_{i,2,A} + \beta_3 ABS_{i,2,A} \times \text{IC}_i \\ + \beta_4 \widehat{\text{Confidence}}_{i,1,A} + \beta_5 \text{IC}_i + \delta_A + \nu_i + \epsilon_{i,A},\end{aligned} \tag{18}$$

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 17 and β_2 in Equation 18). In Equation 18, 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.3 presents the pre-registered test of our hypothesis about experience. We use the same random-effects linear-regression specification as in Equation 10, but include terms interacted with the role that Predictor i played in the Choice Stage:

$$\begin{aligned}\text{Guess}_{i,2,A} = \beta_0 + \beta_1 \text{Guess}_{i,1,A} + \beta_2 S_{i,A} + \beta_3 S_{i,A} \times \text{IC}_i + \beta_4 \text{IC}_i + \beta_5 S_{i,A} \times \text{Exp}_i \\ + \beta_6 S_{i,A} \times \text{IC}_i \times \text{Exp}_i + \beta_7 \text{IC}_i \times \text{Exp}_i + \beta_8 \text{Exp}_i + \beta_9 \bar{S}_{T,A} + \delta_A + \nu_i + \epsilon_{i,A},\end{aligned} \tag{19}$$

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.

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.3.

C Appendix C: Experimental Instructions

C.1 Sentiment-Stage 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.

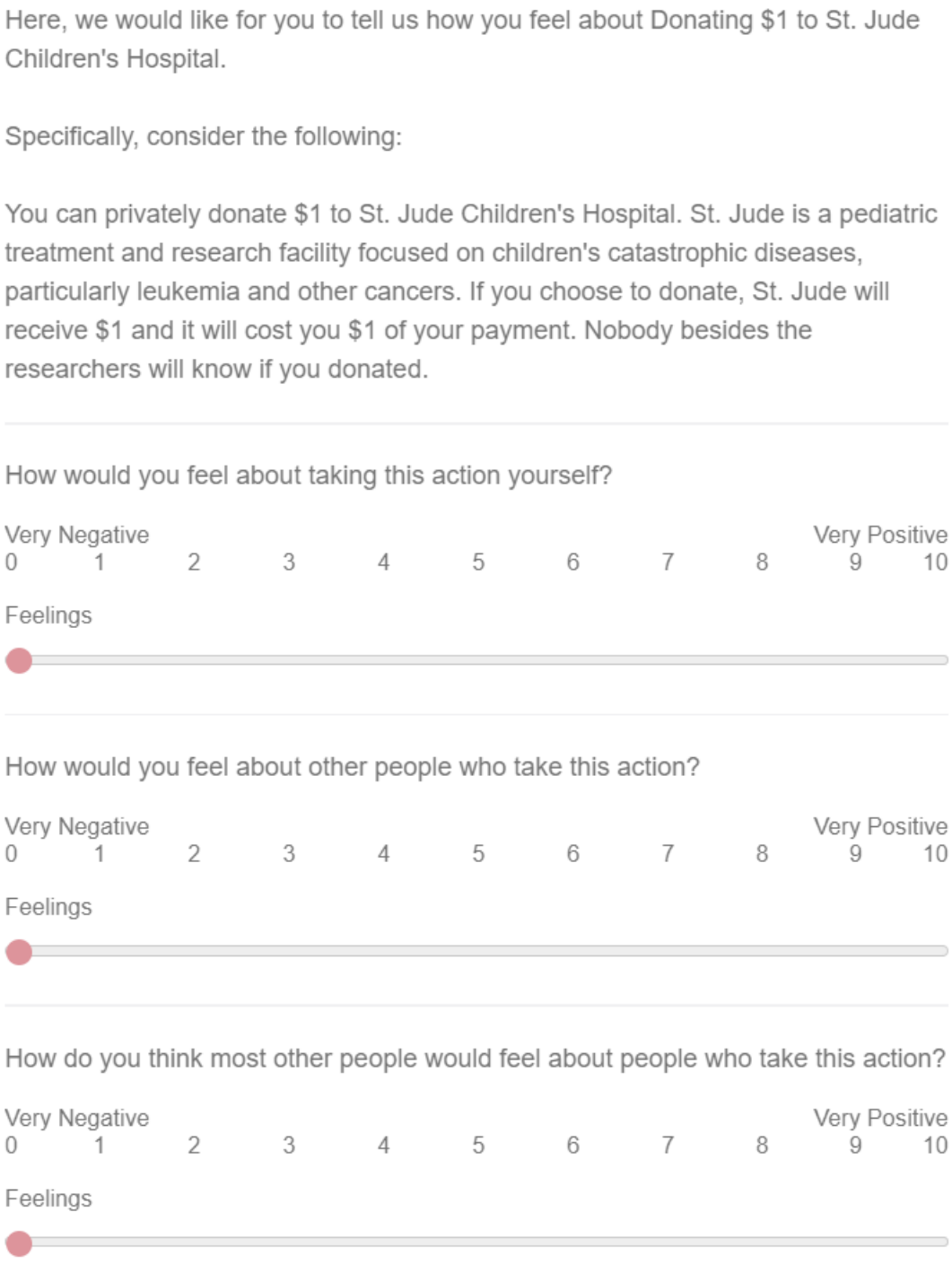
C.1.1 Sentiment-Stage Comprehension Question

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 counts.
- Yes, on will be chosen at randomly-chosen.
- No, you are just asking my opinion.

Figure C.1. Sentiment-Stage Decision Screen



C.2 Choice-Stage 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-Stage 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-Stage 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-Stage 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.

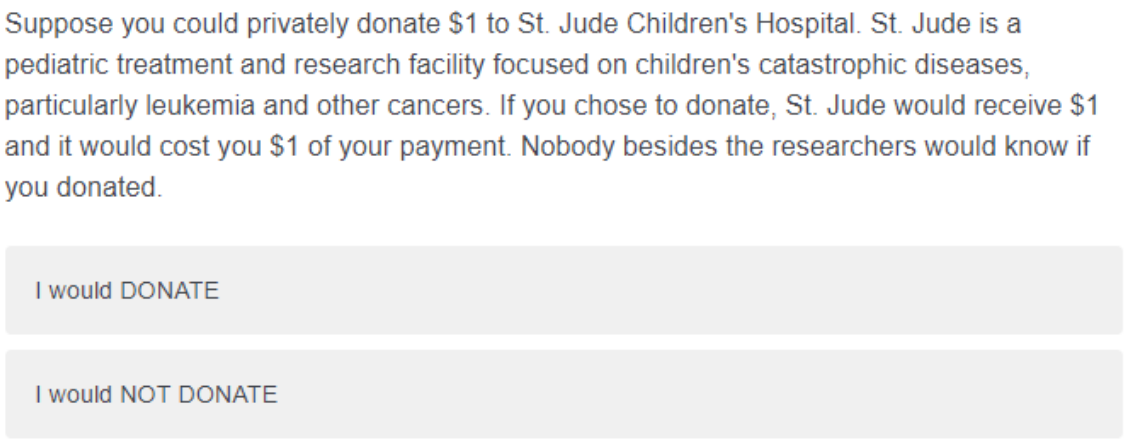
At the end of the study, we will randomly select one of your choices to be the Choice That Counts. The Choice That Counts will determine your outcome today. Since any choice can be selected as the Choice That Counts, you should treat every choice like it is the Choice That Counts.

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?

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

Figure C.2. Hypothetical Decision

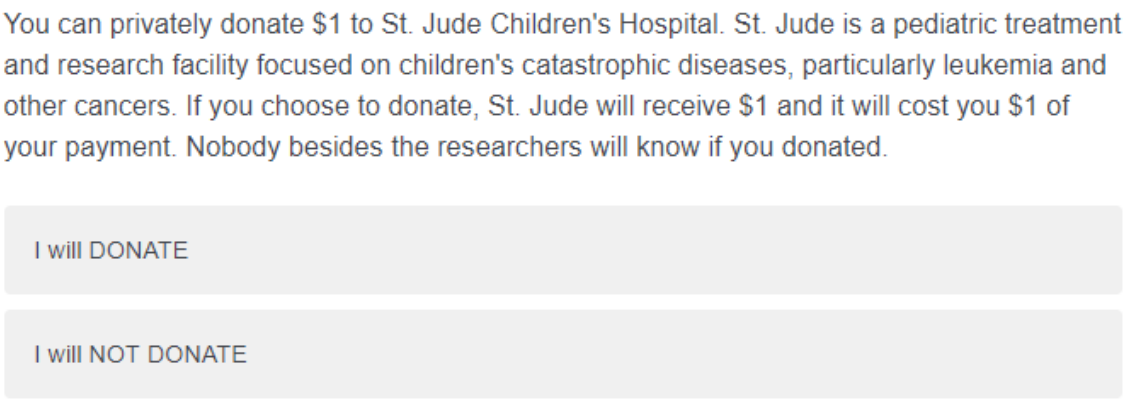


C.4 Prediction-Stage Instructions

C.4.1 Prediction-Stage 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.

Figure C.3. IC Decision



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-Stage 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-Stage 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-Stage 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-Stage 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-Stage 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-Stage 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.

Figure C.4. First-Prediction Choice

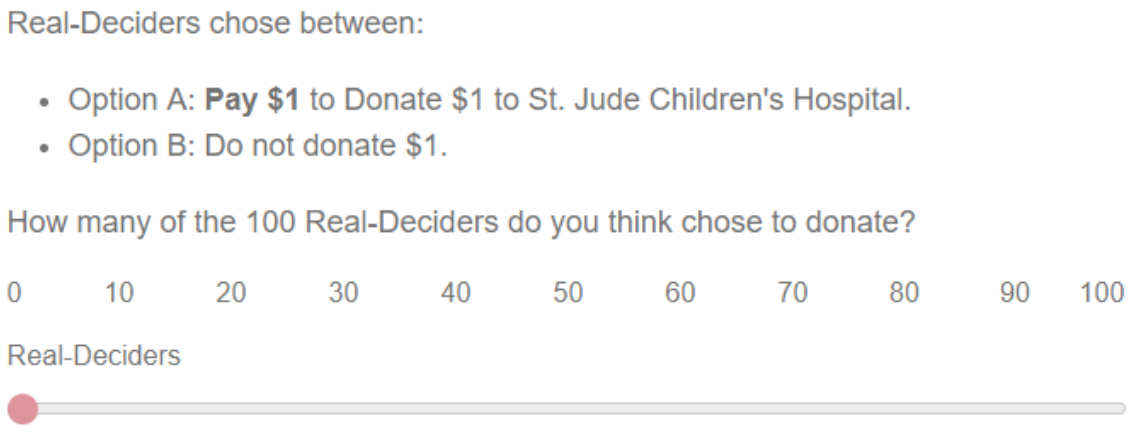
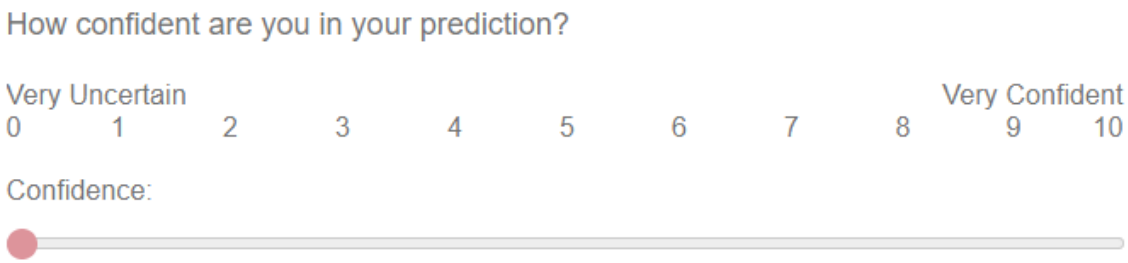


Figure C.5. First-Prediction Choice Confidence



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.

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

Figure C.6. Second Prediction Choice

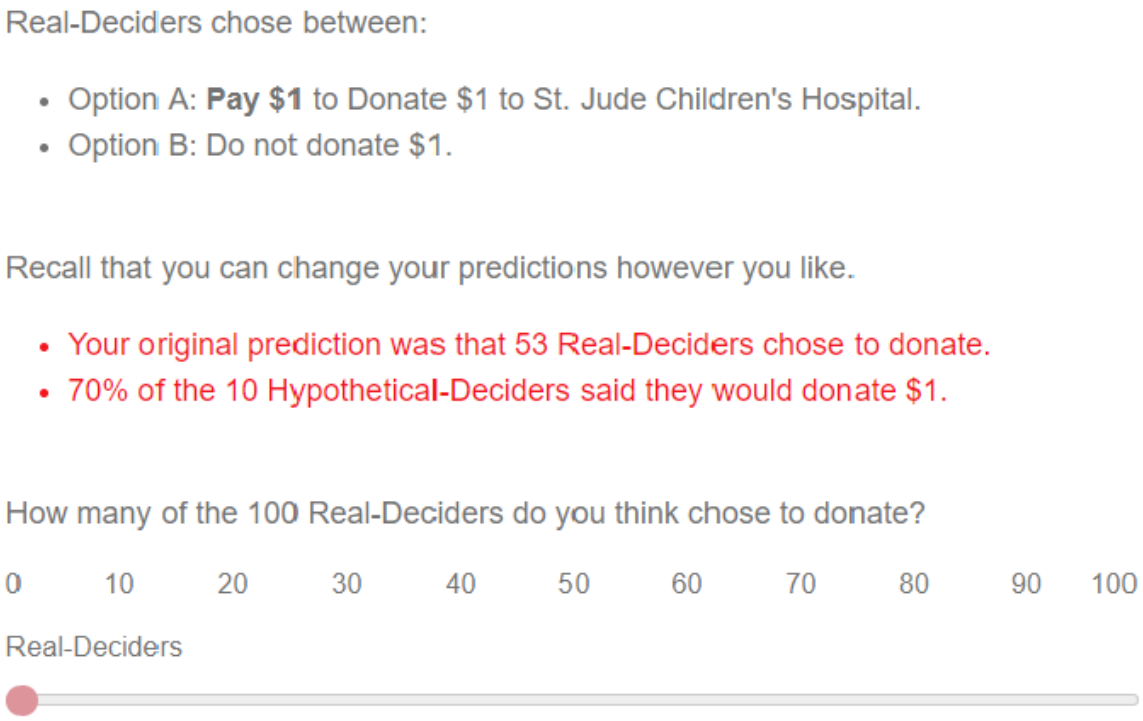


Figure C.7. Second Prediction Choice Confidence