Cognition 256 (2025) 106050

Contents lists available at ScienceDirect

Cognition

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

Disclosing sample bias fails to fully correct judgments of partisan extremity

Alexandra M. van der Valk [a,b,*], Alexander C. Walker [c], Jonathan A. Fugelsang [a], Derek J. Koehler [a]

[a] Department of Psychology, University of Waterloo, Waterloo, ON, Canada
[b] Geriatric Medicine Research, Nova Scotia Health & Dalhousie University, Halifax, NS, Canada
[c] Department of Cognitive & Psychological Sciences, Brown University, Providence, RI, USA

ARTICLE INFO

Keywords:
Social inference
Partisan perceptions
bias
Sample selection
Political polarization
Judgment

ABSTRACT

How do we infer the beliefs of an entire group (e.g., Democrats) after being exposed to the beliefs of only a handful of group members? What if we know that the beliefs we encountered were selected in a biased manner? Across two experiments, we recruited 640 U.S. residents and assessed whether they could recognize and correct for such sample bias. Some participants viewed biased samples that exclusively featured the political opinions of extreme partisans, while others viewed representative samples free from selection biases. Results suggest that people do attempt to correct for known sample bias, but their efforts are often insufficient, leading them to make inaccurate inferences that align with sample bias. Specifically, participants tended to overestimate the ideological extremity of both Democrats and Republicans to a greater extent when exposed to explicitly biased samples, as opposed to representative ones. They also perceived members of the political party in question as holding more homogenous views, presumably because samples of extreme party members' views tend to have less variability than representative samples. Perhaps as a consequence, participants exposed to what they knew to be a biased sample, and who subsequently gave more biased estimates, did not express lower confidence in their estimates compared to participants who were shown representative samples. We discuss how a tendency to insufficiently adjust for transparently biased samples may contribute to partisan misperceptions that fuel political polarization.

1. Introduction

Political opinions are ubiquitous. Politicians and political pundits regularly engage in public discussions on polarizing issues. Meanwhile, social media bombards us with the political attitudes of friends, family, and strangers. We pay attention to what others believe, adjusting our own beliefs in response. For example, telling Republicans that most Republicans agree that the climate is changing increases the likelihood that they themselves endorse this belief (Bayes et al., 2020). How people *perceive* the attitudes of others similarly shapes their viewpoints. Overestimating the ideological extremity of others' political attitudes increases the extremity of one's own (Ahler, 2014). Likewise, exaggerating the degree to which opposing partisans dislike one's political in-group facilitates reciprocal feelings of out-group animosity (Moore-Berg et al., 2020). Fortunately, these misperceptions can be corrected; interventions can reduce both hostility between parties and the extremity of individual political views (Ahler, 2014; Lees & Cikara, 2020).

Despite the prevalence of political expressions, people cannot directly observe the normative beliefs of a group. Rather, such beliefs must be inferred from available evidence. For example, when a Democrat is considering a polarizing topic, they cannot access the complete distribution of beliefs held by other Democrats. Instead they must infer the distribution by piecing together observations of political expressions from various sources like conversations, television, social media, and so on. For the context of this paper, consider this set of observations as a sample of beliefs drawn from a larger population whose overall distribution we wish to infer.

Often, the samples available to us are influenced by biased selection processes, making them less representative of the population from which they are drawn. In the political domain, attention-grabbing polarized content is often overrepresented (Brady et al., 2020; López-Pérez et al., 2022), making more extreme political beliefs seem more

* Corresponding author at: Geriatric Medicine Research, Camp Hill Veteran's Memorial Building, QEII Health Sciences Centre, 1314-5955 Veterans' Memorial Lane, Halifax, NS, B3H 2E1, Canada.
E-mail address: Alexandra.vanderValk@nshealth.ca (A.M. van der Valk).

https://doi.org/10.1016/j.cognition.2024.106050
Received 24 May 2024; Received in revised form 14 December 2024; Accepted 17 December 2024
Available online 2 January 2025

representative of the average partisan. On social media, most political content comes from users who are more politically engaged and ideologically extreme than the average person (Hughes, 2019). There are social rewards for expressing more polarizing views, as evidenced by American politicians with extreme ideological positions gaining more followers on social media compared to their moderate peers (Hong & Kim, 2016), supported by work finding that people are drawn to extreme partisans (Goldenberg et al., 2023). Similarly, using politically-biased language can enhance the perceived trustworthiness of in-group speakers (Walker et al., 2025), while negative tweets about opposing groups often receive higher engagement (Rathje et al., 2021). Rewarding people for expressing polarizing political views helps ensure that the political attitudes people often encounter—whether through social media or other channels like cable news—are more extreme than those held by the average person.

Exposure to biased samples may systematically distort how people perceive the political attitudes of others. People are poor integraters of social judgments and often fail to scrutinize how relationships between network members can bias their perceptions (Fränken et al., 2024). On social media, the content selection processes that amplify particular political expressions are often opaque, concealing biases. Thus, people may fail to recognize that the polarizing views they encounter do not represent the broader attitudes of the population. Consistent with this claim, past research has linked the consumption of political content to less accurate (and more negative) perceptions of opposing partisans (Yudkin et al., 2019). Regardless of their origin, the pervasiveness of partisan misperceptions is well-established. People tend to overestimate not only the ideological extremity of both in- and out-group party members (Levendusky & Malhotra, 2016) but also the extent to which these individuals are politically engaged (Druckman et al., 2022) and belong to stereotypical partisan groups (Ahler & Sood, 2018). Moreover, people tend to overestimate how much the typical partisan dislikes and dehumanizes their political adversaries (Moore-Berg et al., 2020) and underestimate the extent to which they agree with the views of out-group party members (Dorison et al., 2019). These misperceptions have significant consequences, exacerbating partisan animosities and widening ideological rifts (Lees & Cikara, 2021).

Past research has explored the impact of insensitivity to sample bias in various contexts. Ideally, if people were aware of sample bias, they would adjust their inferences about the population in a manner that counteracts that bias. Yet, people seldom have sufficient knowledge of the biases shaping sample selection, which can exacerbate their impact (Koehler & Mercer, 2009). Previous work suggests that people fail to discount evidence even when informed it was selected in a biased manner (Hamill et al., 1980). For example, people shown a sample of judgments from peers are less likely to revise their own estimates when aware that the sample was selected to reflect judgments similar to their own versus being randomly selected (Yaniv et al., 2009). This is supported by work finding that interdependencies between social network members may be able to negatively impact judgment accuracy when people are instructed to incorporate other's beliefs with their own private evidence (Fränken et al., 2024). Other work, however, has found that belief revision is sufficient when sample selection methods are explicit (López-Pérez, Pintér and Sánchez-Mangas, 2022). But in this study, baseline beliefs were made salient, and the judgment in question was of physical aspects of a neutral stimulus. Yet, even when people are informed about how viewing biased samples can skew judgments and are given explicit instructions to refrain from anchoring to presented information, people fail to completely correct for sampling bias, adjusting their judgments insufficiently (Wilson et al., 1996). Thus, encountering biased samples of political opinion may foster partisan misperceptions even when people are explicitly told about the biases involved in sample selection.

We explore this in the current study, presenting an idealized scenario in which people are explicitly informed about the biases shaping the samples they encounter. Across two experiments, participants estimated the average level of agreement among Democrats [Republicans] with various political statements (e.g., "*The US has loose gun laws*") after viewing biased samples of agreement ratings from Democrats [Republicans]. Notably, participants were told about the underlying sampling process that selectively displayed the agreement ratings of more extreme partisans. Our goal was to test how adequately people are able to correct for such sample bias in their estimates of the average level of agreement among all Democrats [Republicans]. We frame our analyses in terms of a simple process model in which the participant uses the (biased) sample mean as a starting point for their estimate of the overall population mean (e.g., the mean level of agreement across all Democrats) and then adjusts from there based on other considerations, in particular the known presence of sample bias. For example, on the 0 (completely disagree) to 100 (completely agree) scale used in both studies, if the participant is shown a sample of Democrats' agreement ratings with a mean of 90, the model assumes that the participant starts with 90 as their initial estimate of the population mean for all Democrats. Because the participant has been told that the sample was drawn from the most extreme Democrats, they would then be expected to adjust their estimate downward toward the middle of the scale to correct for the known sample bias.

Using this model, we define the magnitude of adjustment as the difference between the participant's estimate and the mean of the sample they were shown. In this research we ask (a) whether people do in fact adjust their estimates in an attempt to correct for known sample bias and (b) if so, whether this adjustment is sufficient. There are different standards for assessing sufficiency of adjustment, but the primary one we use is based on a comparison to estimates from an unbiased sample condition. By that standard, adjustment for sample bias is considered sufficient if it brings the final estimates in line with those made in the absence of sample bias. We hypothesized that (a) people would adjust their estimates in an attempt to correct for known sample bias but (b) that such adjustments would typically be insufficient to fully correct for the impact of sample bias.

We recognize that there are other possible process models that could be used to describe how people make estimates in the studies we report below. Likewise, alternative normative frameworks (e.g., Bayesian belief updating) might be applied to determine how a rational agent ought to ideally correct for the impact of sample bias, and in this way adopt different standards against which to define what counts as sufficient adjustment. We sketch a few such possibilities in the general discussion. Our purpose here in introducing the model sketched above is simply to offer it as a means of explicating the measures we use and the hypotheses we test in these studies.

In summary, we hypothesized that viewing biased samples featuring the political attitudes of extreme partisans would be associated with participants overestimating the ideological extremity of the average partisan, even when sample biases were explicitly described. In this way, we aim to investigate one mechanism – specifically, insufficient adjustment from biased samples – that may help to explain why people often misperceive the political attitudes of others, culminating in ways that exacerbate partisan animosities and widen political divides. Experiment 1 provides a test of such insufficient adjustment, while Experiment 2 assesses the extent to which this adjustment reflects a deliberate attempt to correct for known sample bias.

The particular form of sample bias that we consider, in which the most extreme views on one side of the partisan divide are overrepresented, results not only in a sample mean that is more extreme than the population mean, but also in a sample that has less variability from one individual observation to the next within the sample than would be expected in an unbiased sample. We explored two possible additional effects of exposure to the biased sample that could arise as a result of this reduced variability. First, people exposed to the biased sample may perceive greater party consensus (e.g., agreement among Democrats) on the issue compared to those exposed to an unbiased sample in which there is greater variability. Second, although we might generally

anticipate reliance on a biased sample to reduce estimation accuracy (as hypothesized above), when asked to assess the accuracy of their own estimates, any concerns participants have about relying on a sample known to be biased may be offset by the greater coherence (i.e., lower variability) among observations in the sample (e.g., Kahneman & Tversky, 1973).

2. Experiment 1

2.1. Methods

2.1.1. Participants

We recruited 300 United States residents from Amazon Mechanical Turk (MTurk) on October 13th, 2022. To maximize data quality, participants were exclusively recruited from CloudResearch's pool of approved participants (Hauser et al., 2023) and were required to (a) pass two pre-study attention checks and (b) possess an MTurk approval rating equal to or greater than 95 %. We excluded data from 19 participants who failed a post-task comprehension check and one participant who failed to provide sufficient study data, leaving data from 280 participants (60 % Male; 135 Democrats, 66 Republicans, 75 Independents[1]) to be analyzed.

2.1.2. Materials

Experiment 1 featured 24 politically polarizing statements adapted from Vlasceanu et al. (2021). Of these 24 statements, twelve were *Democrat-leaning* statements shown to elicit agreement from Democrats and disagreement from Republicans (e.g., "*The US has loose gun laws*"), while twelve were *Republican-leaning* statements that produced agreement from Republicans and disagreement from Democrats (e.g., "*The US justice system is fair to racial minorities*"). Vlasceanu and colleagues collected a representative sample of 352 Democrats and 352 Republicans, who rated their agreement with each statement on a 101-point scale ranging from 0 (*Completely disagree*) to 100 (*Completely agree*). In our study, we presented participants with the 12 Democrat-leaning [Republican-leaning] statements and asked them to estimate the average rating given by Democrats [Republicans] who participated in the original survey (Vlasceanu et al., 2021; See Fig. 1). Instructions described this survey to participants and informed them whether they would be shown Democrat- or Republican-leaning statements (see Supplementary Materials Part A). On each trial, participants were presented with a statement and instructed to "Please estimate the average agreement rating given by all [Democrat/Republican] participants in the survey." Participants provided these estimates using the aforementioned 101-point scale.

With the exception of those randomly assigned to a No Sample condition, participants were shown the attitudes (i.e., agreement ratings) of five survey respondents on each trial.

No Bias Sample. For each statement, we generated a sample of five agreement ratings that accurately reflected the attitudes of a specific target group (i.e., Democrats or Republicans). To achieve this, for each statement, we divided the ratings of the target respondents into quintiles and then extracted the median rating of each quintile, resulting in five agreement ratings that represented the prevailing attitudes within the target political party (Democrat Item Set: $M_{Respondents} = 69.60$, $M_{Sample} = 69.63$; Republican Item Set: $M_{Respondents} = 63.63$, $M_{Sample} = 63.59$). For example, for each Democrat-leaning statement, we divided the ratings of Democrats into quintiles and displayed the ratings of five survey respondents whose attitudes aligned with the median rating within each quintile (see Fig. 1A). Descriptive statistics of each item are presented in the Supplement Part C.

Prior to the task, participants randomly assigned to view representative samples were informed that the samples they were about to see were chosen in "an unbiased manner" and thus could be regarded as "representative of [Democrats/Republicans] survey respondents as a whole."

Disclosed Bias Sample. Experiment 1 also featured biased samples, wherein the attitudes of five extreme partisans were depicted for each statement. When creating biased samples we considered only the attitudes of Democrats and Republicans whose mean agreement rating across ideologically-congruent statements fell within the top 10 % of respondents. In other words, we considered only the attitudes of Democrats who most strongly endorsed Democrat-leaning statements and Republicans who most strongly endorsed Republican-leaning statements. Mirroring the procedure for generating representative samples, we divided the ratings of these more extreme partisans into quintiles and then extracted the median rating of each quintile. This resulted in agreement ratings from five survey respondents that were representative of the attitudes prevalent among the top 10 % of extreme Democrats or Republicans (Democrat Item Set: $M_{Respondents} = 88.82$, $M_{Sample} = 89.7$; Republican Item Set: $M_{Respondents} = 85.13$, $M_{Sample} = 85.60$), but were unrepresentative (i.e., more extreme and homogenous) of the attitudes of the average partisan respondent. Item-level descriptive statistics of the biased subsample responses are presented in the Supplement Part C. Prior to the task, participants assigned to view biased samples were informed that the samples they were about to see were "selected in a biased manner" and thus could *not* be considered representative of [Democrat/Republican] survey respondents as a whole. Likewise, during each experimental trial, these participants were reminded that the survey respondents presented were "randomly selected from the top 10% of [Democrats/Republicans] who most strongly agreed, on average, with the 12 statements" they would be evaluating (see Fig. 1B).

2.1.3. Measures

Belief Correction. For the two conditions in which participants were presented with a sample of agreement ratings, we calculated the difference between participants' estimates and the mean of the samples they had been shown. A belief correction measure was created by averaging these differences over all 12 statements. Positive [negative] values indicate that a participant's mean estimate was higher [lower] than the mean rating of the presented sampled respondents.

Accuracy. After providing their estimates for all 12 statements, we evaluated participants' perceptions of their accuracy. Our goal was to determine whether participants shown biased samples anticipated being less accurate than those viewing unbiased samples. We defined a "hit" as a statement for which a participant's estimate of the average agreement rating among all Democrats [Republicans] fell within 10 points of the actual mean agreement rating among that group. Following this explanation, participants were asked: "Across your 12 estimates, how many hits do you think you scored?" Participants indicated their subjective accuracy by selecting a number between 0 and 12. Additionally, we calculated an objective accuracy score for each participant that was equal to the number of hits they achieved.

Perceived Party Consensus. Following the task, participants estimated the likelihood (0–100 %) that, for any given issue, two randomly selected Democrats (for those exposed to Democrat-leaning statements) or Republicans (Republican-leaning statements) would provide agreement ratings within 10 points of each other. Given that participants in the Disclosed Bias condition were shown samples that were not only more extreme but also more homogenous than those in the No Bias condition who viewed unbiased samples, we aimed to assess whether they would perceive greater consensus within political groups.

2.1.4. Design and procedure

Experiment 1 employed a 2 × 3 between-subjects design. Participants were presented either 12 Democrat-leaning or 12 Republican-leaning statements and estimated the degree to which the average Democrat or Republican, respectively, agreed with each statement. Additionally, participants were randomly assigned to one of three

[1] Four participants did not report an affiliation.

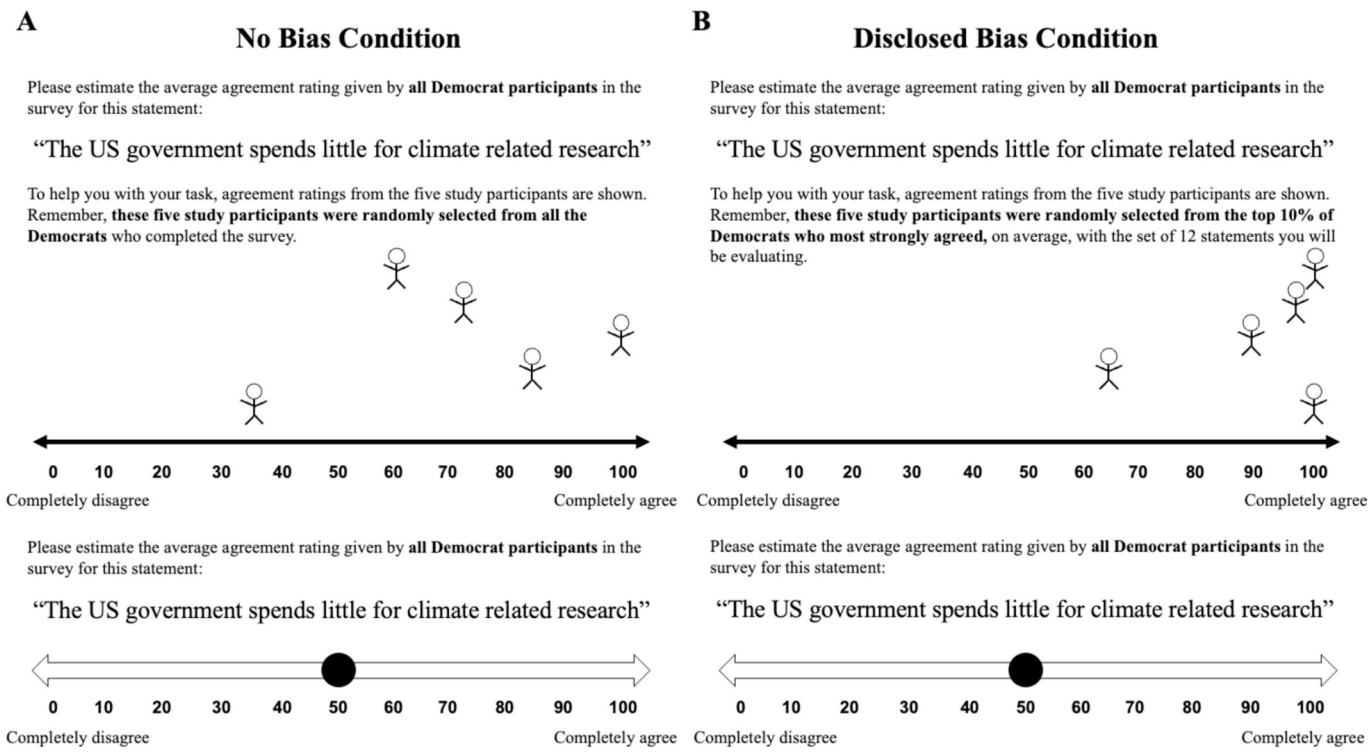

Fig. 1. Example of an item presented to participants in No Bias (Panel A) and Disclosed Bias (Panel B) conditions.

conditions: providing each estimate without sample information (*No Sample*), after viewing a representative sample of five survey respondents (*No Bias*), or after viewing a biased sample of five ideologically extreme respondents (*Disclosed Bias*). Participants in the representative or biased sample conditions were informed about the relevant sample selection method and completed pre- and post-task comprehension checks. Following experimental trials, participants responded to accuracy and perceived consensus questions and provided demographic information.

2.2. Results

Mean agreement estimates—collapsed across Democrat- and Republican-leaning statements—in the Disclosed Bias, No Bias, and No Sample (control) conditions are shown in Fig. 2. First, we ask if participants shown biased samples adjust their estimates in attempt to correct for sample biases, by comparing their belief correction scores to those of participants in the No Bias condition. Consistent with such adjustment, belief correction was more pronounced in the Disclosed Bias ($M = -12.94$, $SD = 12.53$) compared to the No Bias condition ($M = 2.59$, $SD = 7.55$), $t(179) = 9.65$, $p < .001$, $d = 1.45$, 95 % *CI* [1.12, 1.78], with participants in the Disclosed Bias condition providing less extreme estimates than the mean of the biased samples they were shown, $t(103) = 11.10$, $p < .001$, $d = 1.09$, 95 % *CI* [0.84, 1.33].

Second, we assess the sufficiency of adjustment in the Disclosed Bias condition by comparing the mean estimate in this condition to that in the No Bias condition (and also the control, No Sample condition). Although all participants overestimated the ideological extremity of the average partisan ($p's < 0.001$, $d's > 0.45$), this tendency was most pronounced among participants in the Disclosed Bias condition, suggesting

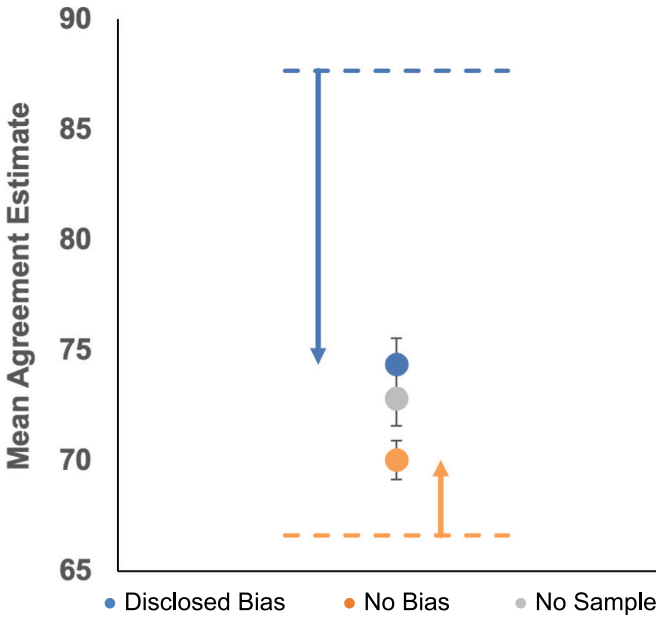

Fig. 2. Experiment 1: Mean Agreement Estimates of the Average Partisan. Dashed lines represent the mean agreement ratings of survey respondents presented in biased (blue) and unbiased (orange) samples. Arrows depict the difference between sample means and participants' mean agreement rating estimates of the average partisan within a condition. Error bars represent +/− 1 SE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a failure to fully correct for sample bias, despite such bias being disclosed. A one-way analysis of variance (ANOVA) revealed that estimated agreement ratings of the average partisan differed based on Sample Type (Disclosed Bias, No Bias, No Sample), $F(2, 277) = 3.26$, $p = .040$, $\eta_p^2 = 0.023$.[2] One-tailed independent samples *t*-tests indicated that estimates were more extreme in the Disclosed Bias ($M = 74.33$, $SD = 12.25$) as opposed to the No Bias condition ($M = 70.03$, $SD = 7.60$), $t(179) = 2.72$, $p = .004$, $d = 0.41$, 95 % *CI* [0.11, 0.71]. However, while estimates in the No Bias condition were slightly more moderate than those in the No Sample condition ($M = 72.82$, $SD = 12.48$), $t(174) = 1.73$, $p = .043$, $d = 0.26$, 95 % *CI* [0.04, 0.56], estimates did not reliably differ between the Disclosed Biased and No Sample conditions, $t(201) = 0.87$, $p = .192$, $d = 0.12$, 95 % *CI* [−0.15, 0.40]. Factorial ANOVAs exploring the effects of participant party affiliation and affiliation strength (along with Sample Type) showed no significant effects of these variables (*p's* > 0.344), indicating that agreement estimates did not vary based on participants' political stance (see Supplementary Materials Part B).[3]

2.2.1. Accuracy

Subjective and objective accuracy scores are illustrated in Fig. 3. A one-way ANOVA revealed that subjective accuracy did not vary based on Sample Type, $F(2, 277) = 0.74$, $p = .480$, $\eta_p^2 = 0.005$. Participants shown biased samples did not expect their estimates to be any less accurate than did participants in the No Bias and No Sample conditions. However, *objective* accuracy reliably differed by Sample Type, $F(2, 277) = 59.01$, $p < .001$, $\eta_p^2 = 0.299$. Participants in the No Sample ($M = 3.62$, $SD = 2.31$) and Disclosed Bias ($M = 4.29$, $SD = 3.18$) conditions were similarly inaccurate in their agreement rating estimates of the average partisan ($p = .087$, $d = 0.24$), and were considerably less accurate than those in the No Bias condition ($M = 8.25$, $SD = 3.44$; *p's* < 0.001, *d's* > 1.20). Taken together, participants viewing representative samples of political attitudes (No Bias) slightly underestimated the accuracy of their estimates, $t(76) = 3.07$, $p = .003$, $d = 0.35$, 95 % *CI* [0.12, 0.58], while those viewing no sample information or transparently biased sample information were considerably overconfident (*p's* < 0.001, *d's* > 0.62).

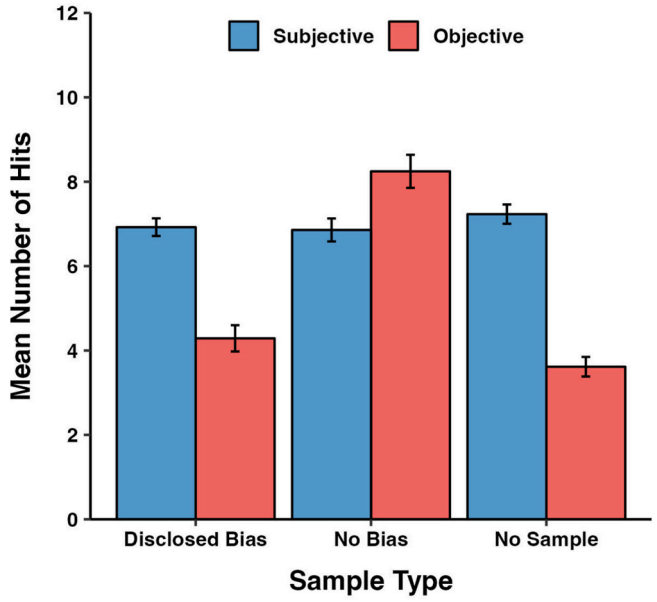

Fig. 3. Experiment 1: Subjective and Objective Accuracy. Bars display participants' mean subjective and objective accuracy scores within the Sample Type condition for which participants were randomly assigned. Error bars represent +/− 1 SE.

2.2.2. Perceived party consensus

Perceptions of within-party consensus differed by Sample Type, $F(2, 277) = 4.18$, $p = .016$, $\eta_p^2 = 0.029$. Participants viewing biased samples (which tended to have less variance) in the Disclosed Bias condition perceived greater consensus ($M = 68.22$, $SD = 19.33$) among the attitudes of Democrats and Republicans compared to those viewing unbiased samples in the No Bias condition ($M = 61.10$, $SD = 16.73$), $t(179) = 2.59$, $p = .010$, $d = 0.39$, 95 % *CI* [0.09, 0.69]. Perceptions of party consensus did not, however, differ between Disclosed Bias and No Sample conditions ($M = 68.42$, $SD = 19.35$), $t(201) = 0.08$, $p = .940$, $d = 0.01$, 95 % *CI* [−0.27, 0.29].

3. Experiment 2

The results of Experiment 1 suggest that participants exposed to biased samples deliberately adjust their inferences to compensate for sample bias. This could explain why their estimates deviate further from the mean of the samples they viewed compared to people shown unbiased samples. There is, however, a plausible alternative explanation for these results. Participants may have disregarded the instructions indicating whether their samples were biased or unbiased, instead treating both conditions as if they were representative samples. From a Bayesian perspective, participants may have simply adjusted from a prior toward the sample mean they were shown. If their prior fell closer to the unbiased sample mean than the biased sample mean – a reasonable assumption – this could produce a qualitative pattern of results akin to those observed in Experiment 1. Hence, Experiment 1 does not conclusively demonstrate that participants viewing biased samples deliberately corrected for known sample bias. Experiment 2 addresses this potential limitation by replacing the No Sample condition with another biased sample condition, where participants were *not* informed that the presented samples were biased (Undisclosed Bias condition). Greater belief correction in the Disclosed Bias condition compared to the Undisclosed Bias condition would suggest that participants deliberately adjust their estimates in response to known sample bias.

3.1. Methods

3.1.1. Participants

Three hundred and forty US residents were recruited from MTurk on May 23rd, 2023 using the same recruitment criteria as Experiment 1. Those who participated in Experiment 1 were restricted from participating in Experiment 2. We excluded data from 45 participants who failed a post-task comprehension check, leaving data from 286 participants (56 % Male; 165 Democrats, 64 Republicans, 55 Independents[4]) to be analyzed.

3.1.2. Materials and measures

Experiment 2 retained the same materials and measures as Experiment 1 with one modification: we replaced the No Sample condition

[2] Across studies, the target party (Democrat or Republican) that participants were randomly assigned to evaluate exerted at most a minimal impact on their judgments. As such, all analyses including Target Party are exclusively reported in the supplementary materials (Part B).

[3] As is common in online crowdsource data, a lower proportion of self-reported Republicans opted to participate in both experiments. Though we found no effect of participant political affiliation on final estimates, this characteristic should be heeded. We report these null findings for both experiments in the Supplement Part B.

[4] Two participants did not report an affiliation.

from in Experiment 1 with a new Undisclosed Bias condition. Participants assigned to the Undisclosed Bias condition were exposed the same biased samples as those in the Disclosed Bias condition. However, unlike participants in the Disclosed Bias condition, those in the Undisclosed Bias condition were falsely told that the ratings of the sampled survey respondents were randomly selected from all Democrat [Republican] respondents. Thus, although the samples presented in the Undisclosed Bias condition matched those in the Disclosed Bias condition, the instructions provided in the Undisclosed Bias condition mirrored those given in the No Bias condition.

3.1.3. Design and procedure

The design and procedure of Experiment 2 matched that of Experiment 1. Participants estimated the degree to which either the average Democrat or Republican agreed with 12 Democrat- or 12 Republican-leaning statements, respectively. Based on random assignment, participants provided each estimate after viewing a representative sample of five respondents (No Bias condition), viewing an explicitly biased sample of five ideologically extreme respondents (Disclosed Bias condition), or viewing a biased sample of five extreme respondents falsely depicted as being representative (Undisclosed Bias condition). Similar to Experiment 1, participants responded to accuracy and perceived consensus questions following experimental trials[5] and provided demographic information prior to the post-study debriefing.

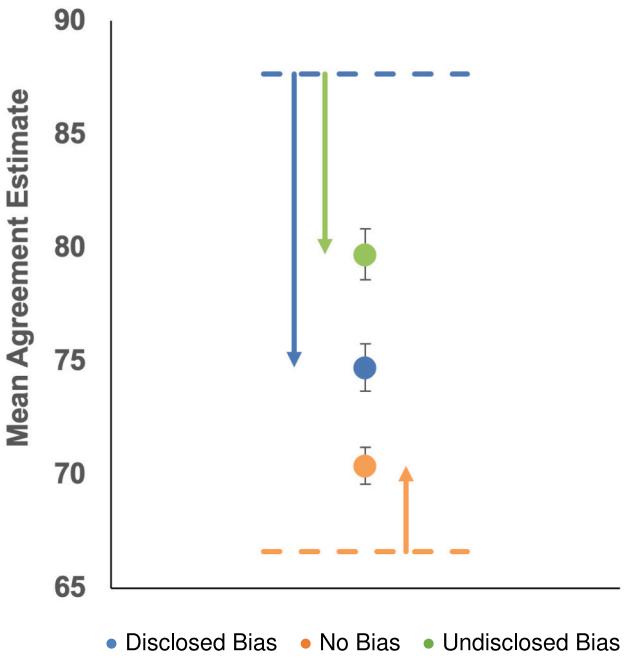

Fig. 4. Experiment 2: Mean Agreement Rating Estimates of the Average Partisan. Dashed lines represent the mean agreement ratings of survey respondents presented in biased (blue/green) and unbiased (orange) samples. Arrows depict the difference between the relevant sample mean and participants' mean agreement rating estimates of the average partisan within a condition. Error bars represent +/− 1 SE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Results

Mean estimates from the Disclosed Bias, No Bias, and Undisclosed Bias conditions are shown in Fig. 4. Belief correction varied across Sample Type, $F(2, 283) = 72.70$, $p < .001$, $\eta_p^2 = 0.339$. In the critical comparison, belief correction was greater in the Disclosed Bias condition ($M = -12.78$, $SD = 10.25$) than in the Undisclosed Bias condition ($M = -7.86$, $SD = 10.66$), $t(190) = 3.26$, $p = .001$, $d = 0.47$, 95 % *CI* [0.18, 0.76], suggesting deliberate adjustment in response to known sample bias. Belief correction was also greater in the Disclosed Bias compared to the No Bias condition ($M = 3.73$, $SD = 7.90$), $t(189) = 12.44$, $p < .001$, $d = 1.80$, 95 % *CI* [1.46, 2.14]. As in Experiment 1, adjustment for known sample bias was insufficient. Critically, participants in the Disclosed Bias condition perceived the average partisan as agreeing more strongly with ideologically-congruent statements ($M = 74.73$, $SD = 10.32$) than those in the No Bias condition ($M = 70.40$, $SD = 7.86$), $t(189) = 3.25$, $p = .001$, $d = 0.47$, 95 % *CI* [0.18, 0.76]. Furthermore, participants in the Disclosed Bias condition provided more moderate estimates than those in the Undisclosed Bias condition ($M = 79.71$, $SD = 10.91$), $t(190) = 3.25$, $p = .001$, $d = 0.47$, 95 % *CI* [0.18, 0.76]. Therefore, viewing biased—as opposed to representative—samples of political attitudes increased partisan misperceptions, leading people to significantly overestimate the ideological extremity of the average partisan,[6] while transparency regarding the biases inherent in sample selection helped lessen the severity of these misperceptions.

3.2.1. Accuracy

As in Experiment 1, participants in the Disclosed Bias ($M = 7.43$, $SD = 2.19$) and No Bias conditions ($M = 7.36$, $SD = 2.20$) perceived their agreement rating estimates to be similarly accurate, $t(189) = 0.23$, $p = .823$, $d = 0.03$, 95 % *CI* [−0.25, 0.32], despite those in the No Bias condition being objectively more accurate in their judgments, $t(189) = 6.43$, $p < .001$, $d = 0.93$, 95 % *CI* [0.63, 1.23]. Thus, while the subjective accuracy of participants in the No Bias condition was well calibrated, those viewing biased samples were again considerably overconfident, particularly when the biases inherent in sample selection were undisclosed (see Fig. 5).

3.2.2. Perceived party consensus

Perceptions of within-party consensus differed by Sample Type, $F(2, 282) = 4.05$, $p = .019$, $\eta_p^2 = 0.028$. Participants exposed to biased (and more homogenous) samples perceived greater consensus among the attitudes of Democrats and Republicans compared to those shown unbiased samples. Specifically, participants in the Undisclosed Bias condition perceived greater consensus ($M = 72.73$, $SD = 15.88$) than those in the No Bias condition ($M = 65.42$, $SD = 20.83$), $t(186) = 2.71$, $p = .007$, $d = 0.40$, 95 % *CI* [0.11, 0.68]. Participants in the Disclosed Bias condition also perceived greater consensus ($M = 69.20$, $SD = 15.81$) than those in the No Bias condition, however, this difference was small and not statistically significant, $t(189) = 1.42$, $p = .158$, $d = 0.20$, 95 % *CI* [−0.08, 0.49].

4. General discussion

The tendency for people to overestimate the ideological extremity of the average partisan is a well-documented phenomenon (Ahler, 2014; Westfall et al., 2015), as is the link between partisan misperceptions and political polarization (Lees & Cikara, 2021). However, less is known about *why* people overestimate the extremity of political in- and outgroup members. One explanation is that since directly observing the

[5] Participants also completed four items designed to measure individual differences in numeracy (Cokely et al., 2012). Analyses featuring data from this measure can be viewed in the supplementary materials (Part B).

[6] Notably, regardless of the type of sample shown, participants tended to overestimate the ideological extremity of the average partisan (p's < 0.001, d's > 0.48).

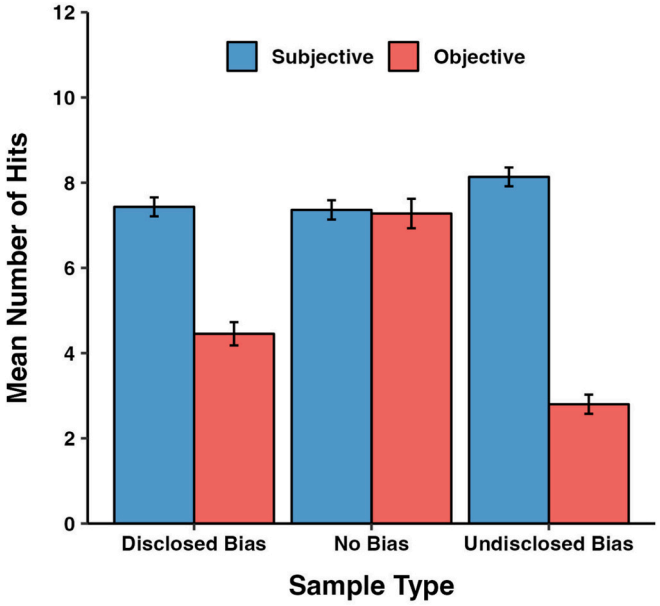

Fig. 5. Experiment 2: Subjective and Objective Accuracy. Bars display participants' mean subjective and objective accuracy scores within the Sample Type condition for which participants were randomly assigned. Error bars represent +/− 1 SE.

complete distribution of political beliefs across a population is unfeasible, people draw inferences about partisan groups based on a sample of expressions they encounter. However, certain beliefs – particularly those held by politically engaged individuals with ideologically extreme views – are overrepresented on social media and other platforms. It stands to reason that the overrepresentation of unrepresentative viewpoints could systematically distort people's perceptions of what the average partisan believes. Consistent with this claim, we observed that people exposed to biased—as opposed to representative—samples of political opinion were more likely to overestimate the ideological extremity of both Democrats and Republicans. Thus, we demonstrate that the political attitudes individuals encounter may exert a sizeable impact on how they perceive the political beliefs of others.

The present work created an idealized scenario in which participants were explicitly informed about the biases shaping the samples they encountered. In both experiments, participants correctly estimated that the beliefs of the average partisan were more moderate than the beliefs depicted in the ideologically extreme samples viewed. Nevertheless, their estimates were more extreme than both the true sample mean and the estimates of people exposed to unbiased samples, revealing that this adjustment was generally insufficient. Notably, estimates of group belief were more accurate when participants were informed of the biases inherent in sample selection, suggesting that they intentionally attempted to correct for sample bias. However, the specific processes responsible for sample bias are seldom transparent. Thus, when sample bias is concealed, as was the case in the Undisclosed Bias condition of Experiment 2, biased samples distorted perceptions of group beliefs such that partisan perceptions were made considerably less accurate, and more ideologically extreme. Our findings are partially supported by a recent study on the neglect of sample selection methods (López-Pérez et al., 2022). The authors instructed participants to estimate the probability that a randomly-selected ball was a specific colour after viewing a subsample of 5 balls. When the biased sample selection method was made transparent, people revised their judgments to be more accurate. When this selection method was partially disclosed – introducing some ambiguity – they (incorrectly) utilized the sample as representative evidence. In slight contrast to our study, their making sample bias transparent led to their participants making near-perfect judgments, whereas in our study, participants improved accuracy greatly yet still significantly overestimated true population beliefs. We opted not to make individual beliefs salient, given we wanted to encourage participants to focus on their perceptions of (the distribution of) the beliefs of others. Furthermore, the stimuli used in López-Pérez et al. resulted in a "low stakes" task. The relevance of the stimuli to the decision maker is useful to probe when studying the impact of sample bias on judgments – in the present study, the dynamic political landscape is highly relevant and familiar to participants.

We might have expected that participants informed of the biased nature of the samples would express less confidence in the accuracy of their estimates compared to those presented with unbiased samples. However, participants in the Disclosed Bias condition did not exhibit a reduction in confidence. It is possible that awareness of sample bias did reduce confidence, but that this impact was offset by the greater homogeneity among observations (which may have increased confidence in one's estimate). The biased samples used in this study were sourced from only the most extreme party members, whose responses agreed with each other more than those featured in representative samples of party members. Consistent with the increased homogeneity of biased samples, participants exposed to sample bias not only overestimated the extremity of Democrats' and Republicans' political opinions but also the degree of consensus within each political party.

Previous research investigating predictions derived from one-sided evidence, such as seeing only the defendant's arguments in a legal case without those of the plaintiff, demonstrated a closely related phenomenon. Despite predictions based on one-sided evidence being less accurate than those based on two-sided evidence, they were not made with lower confidence. Presumably, one-sided evidence presents a more coherent account, which enhances confidence (Brenner et al., 1996).

Taken together, the present work reveals one potential mechanism explaining how biased samples of political expressions, such as those commonly encountered online (Brady et al., 2023), lead people to misperceive the political beliefs of others. Hidden biases in sample selection, such as those that amplify the voices of extreme partisans, can foster partisan misperceptions as individuals fail to realize that these opinions do not represent the group. Even when aware of sample bias, people fail to adequately correct for it, resulting in false perceptions that are aligned with sample biases. Thus, even when transparent, biases in sample selection can facilitate false perceptions as people recognize the need to account for sample bias but fail to adjust their perceptions adequately. For instance, in the political domain, people may recognize that the beliefs of the average person tend to be more moderate than the viewpoints they frequently encounter, yet still fail to appreciate the extent to which these views are unrepresentative of those endorsed by the average person. This isn't to say that holding these misinformed views is wholly detrimental. It could be beneficial, for example, to give greater weight to the most extreme views when predicting group behaviour (e.g. this subset may have an outsize impact on political events, such as a protest initiated by the most extreme members of a political party). But generally, holding misinformed or stereotyped views of a political population presents unique challenges considering how salient partisanship has become in our social world, particularly in the United States (Iyengar & Krupenkin, 2018). Overweighting the most extreme partisans could lead to worse outcomes in scenarios such as trying to disseminate political information to a broad audience, or a politician trying to secure votes – here, optimal results will come from accurately perceiving the beliefs of the target demographic. The ideal scenario would be to overweight extreme members when they are likely to have outsized influence, but focus more on representativeness when such extremity is less relevant.

We have framed the results of our studies using a simple process model in which people use the sample mean as a starting point for their estimate of the overall population mean and then adjust from there based on other considerations, with particular focus on how they adjust for the known presence of sample bias. While we find this model useful

for framing our hypotheses and analyses, we acknowledge that other process models could accommodate our results equally well and, in this way, offer alternative or co-occurring explanations for them. For instance, our model assumes that people take the sample mean as their initial estimate, which seems plausible as judgments are often highly responsive to overtly-presented information (Kahneman, 2011) and tend to integrate others' judgments without critical scrutiny (Fränken et al., 2024). However, it is also plausible that people form an initial estimate based on their prior beliefs and then adjust from there in the direction implied by the sample mean. Even the notion that there is a single starting point for the estimate and some kind of adjustment process that then follows could be questioned (Epley & Gilovich, 2001, 2006). For instance, perhaps people form two separate estimates, one from the sample and one from their prior beliefs, and average – or otherwise combine – the two through calculation. As we did not collect participant priors, we are unable to provide supportive or refutive evidence for such frameworks outside the scope of our study design. Nevertheless, these questions are worthy of future exploration.

The main conclusion we draw, framed in terms of our preferred process model, is that adjustment (as defined within this model) for the presence of sample bias is insufficient to fully correct for its influence. We believe the standard(s) that we adopt in our analyses to define what would count as "sufficient" adjustment are not entirely dependent on the process model we have sketched. That is, basing claims about insufficient adjustment on the comparison of estimates based on biased samples to those based on unbiased samples seems justifiable to us in a way that transcends the specifics of different process models that could be developed to account for our results. That said, we also acknowledge that alternative normative frameworks could be applied to create other standards. For instance, a Bayesian framework could be used to specify how a rational agent should correct for the impact of the form of sample bias investigated in our studies, and application of such a framework might yield different conclusions about whether participants in our studies made "sufficient" adjustment for sample bias.

5. Limitations

The results observed in these experiments should be interpreted minding the following limitation. The data used to compute the real-world samples (sourced from Vlasceanu et al., 2021) were collected in Fall 2019, while our participants provided their estimates of agreement in Fall 2022 (Experiment 1) and Spring 2023 (Experiment 2). While a benefit of this study is its reliance on real-world partisan opinions, these attitudes may have shifted in the past three years. While this may explain participants' overestimations of partisan agreement regarding ingroup statements, it does not explain how these estimates were impacted by the samples viewed (e.g., participants providing more extreme estimates in the Disclosed Bias compared to No Bias condition). Reducing the timespan between when partisan beliefs are collected and when participants make their judgments is a welcome avenue for future investigation. Furthermore, much work suggests that people consistently overestimate the ideological extremity of their political ingroup and outgroup (Homola et al., 2023; Lees & Cikara, 2021; Moore-Berg et al., 2020). Thus, participants overestimating the average partisan's agreement with politically congruent statements in Experiments 1 and 2 may result from a general tendency to perceive the average partisan as more extreme than they are as opposed to indicating an increase in partisan extremity across time periods. This is a fruitful avenue of future exploration.

6. Conclusion

How people perceive the attitudes of others shapes their own political viewpoints, while overestimating the ideological extremity of opposing partisans hinders productive cross-party interactions and amplifies partisan animosities (Ahler, 2014; Lees & Cikara, 2021; Wilson et al., 2020). Accordingly, misattributing the beliefs of the most extreme ideologues to the average partisan can lead people to view rival partisans as holding irreconcilable views, deepening political divides. Therefore, exposure to biased samples, including those resulting from the overrepresentation of ideologically extreme voices (Hughes, 2019), can contribute to increasing political polarization.

Scholars have noted the importance of making content algorithms (e.g., that amplify polarizing content) more transparent, suggesting that better transparency could mitigate social misperceptions by allowing individuals to adjust their inferences to account for biases in content selection (Brady et al., 2023; López-Pérez et al., 2022). Mirroring such a scenario in which these biases are well-described, the present work demonstrates this potential benefit of algorithm transparency. We observed that individuals adjust their inferences to account for known sample bias in a manner that mitigates partisan misperceptions. We concluded, however, that this adjustment is often insufficient. Thus, while algorithm transparency may reduce social misperceptions, such misperceptions are likely to persist as a result of individuals exposure to biased samples. Additional interventions in tandem with algorithm transparency may be prudent. For example, in this work, eliminating sample bias altogether promoted more accurate inferences than did making sample bias explicit. Nevertheless, the ability of polarizing content to capture attention and boost audience engagement (Brady et al., 2020) helps ensure its overrepresentation across platforms. Therefore, understanding how people interact with biased samples—transparent or not—to make social inferences is a worthwhile goal, as inaccurate perceptions not only facilitate inaccurate beliefs but also promote inter-group conflict.

Ethics and consent

This research complies with the Declaration of Helsinki (2013). All experiments were reviewed and received ethics clearance from a University of Waterloo Research Ethics Committee.

Funding

This research was supported by a Social Sciences and Humanities Research Council of Canada Banting Postdoctoral Fellowship (to AW) and by grants from The Natural Sciences and Engineering Research Council of Canada (to JF and DK).

CRediT authorship contribution statement

Alexandra M. van der Valk: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alexander C. Walker:** Writing – review & editing, Visualization. **Jonathan A. Fugelsang:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Derek J. Koehler:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors have no competing interests to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.106050.

Data availability

The data and analysis scripts of all Experiments can be accessed on the Open Science Framework (https://osf.io/g9tmz/). All study mate-

rials can be viewed in the supplementary materials (Part A).

References

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *The Journal of Politics, 76*(3), 607–620.

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics, 80*(3), 964–981.

Bayes, R., Druckman, J. N., Goods, A., & Molden, D. C. (2020). When and how different motives can drive motivated political reasoning. *Political Psychology, 41*(5), 1031–1052.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science, 15*(4), 978–1010.

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences, 27*(10), 947–960.

Brenner, L. A., Koehler, D. J., & Tversky, A. (1996). On the evaluation of one-sided evidence. *Journal of Behavioral Decision Making, 9*(1), 59–70.

Dorison, C. A., Minson, J. A., & Rogers, T. (2019). Selective exposure partly relies on faulty affective forecasts. *Cognition, 188*, 98–107.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2022). (Mis) estimating affective polarization. *The Journal of Politics, 84*(2), 1106–1117.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*(5), 391–396.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311–318.

Fränken, J. P., Valentin, S., Lucas, C. G., & Bramley, N. R. (2024). Naïve information aggregation in human social learning. *Cognition, 242*, Article 105633.

Goldenberg, A., Abruzzo, J. M., Huang, Z., et al. (2023). Homophily and acrophily as drivers of political segregation. *Nature Human Behaviour, 7*, 219–230.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology, 39*(4), 578–589.

Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2023). Evaluating CloudResearch's approved group as a solution for problematic data quality on MTurk. *Behavior Research Methods, 55*(8), 3953–3964.

Homola, J., Rogowski, J. C., Sinclair, B., et al. (2023). Through the ideology of the beholder: How ideology shapes perceptions of partisan groups. *Political Science Research and Methods, 11*, 275–292.

Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly, 33*(4), 777–782.

Hughes, A. (2019, October 23). *A small group of prolific users account for a majority of political tweets sent by U.S. adults*. Pew Research Center. https://pewrsr.ch/3a31iDK.

Iyengar, S., & Krupenkin, M. (2018). Partisanship as social identity; implications for the study of party polarization. *The Forum, 16*(1), 23–45.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237.

Koehler, J. J., & Mercer, M. (2009). Selection neglect in mutual fund advertisements. *Management Science, 55*(7), 1107–1121.

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour, 4*(3), 279–286.

Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B, 376*(1822), 20200143.

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly, 80*(S1), 378–391.

López-Pérez, R., Pintér, Á., & Sánchez-Mangas, R. (2022). Some conditions (not) affecting selection neglect: Evidence from the lab. *Journal of Economic Behavior & Organization, 195*, 140–157.

Moore-Berg, S. L., Ankori-Karlinsky, L. O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences, 117*(26), 14864–14872.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26), Article e2024292118.

Vlasceanu, M., Morais, M. J., & Coman, A. (2021). The effect of prediction error on belief update across the political spectrum. *Psychological Science, 32*(6), 916–933.

Walker, A. C., Fugelsang, J. A., & Koehler, D. J. (2025). Partisan language in a polarized world: In-group language provides reputational benefits to speakers while polarizing audiences. *Cognition, 254*, Article 106012.

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science, 10*(2), 145–158.

Wilson, A. E., Parker, V. A., & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences, 34*, 223–228.

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology. General, 125*(4), 387–402.

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 558.

Yudkin, D., Hawkins, S., & Dixon, T. (2019, June). The perception gap: How false impressions are pulling Americans apart. More in Common. https://perceptiongap.us.