People Believe If 90% Prefer A over B, A Must Be Much Better than B. Are They Wrong?

Graham Overton

Joachim Vosgerau

Ioannis Evangelidis

Forthcoming, *Journal of Consumer Research*

Graham Overton (graham.overton@nus.edu.sg) is a visiting assistant professor of marketing at the National University of Singapore Business School. Joachim Vosgerau (joachim.vosgerau@unibocconi.it) is a professor of marketing at Università Bocconi, Via Röntgen 1, 20136 Milano, Italia. Ioannis Evangelidis (ioannis.evangelidis@esade.edu) is an associate professor of marketing at ESADE, Universitat Ramon Llull, Sant Cugat del Valles, Barcelona, Spain. Please address correspondence to Graham Overton. This manuscript is based on the lead author's dissertation completed under the supervision of the second and third authors. This research was partly supported by Grant PID2021-128099NA-I00 funded by MICIU/AEI/10.13039/501100011033 and FEDER, UE. Supplementary materials are included in the web appendix accompanying the online version of this article.

Editor: Andrew T. Stephen

Associate Editor: Simon J. Blanchard

ABSTRACT

We show that consumers confuse consensus information in polls—such as 90% prefer product A over product B—with differences in liking—the extent to which poll respondents like A better than B. Consequently, they interpret a 90% consensus in favor of A as the average liking of A being considerably higher than the average liking of B. We demonstrate empirically and with simulations that—while this can be true—it is more probable that the average liking of A is only slightly higher than that of B. This regularity is robust to the sign and size of the correlation between ratings for A and B, and across most distributions for A and B's liking. Consumers are not aware of this regularity, and believe that 90% consensus implies A being *much better* than B. Communicators (marketers, managers, public policy makers, etc.) can capitalize on these erroneous inferences and strategically display preference information as consensus or as liking ratings leading to dramatic shifts in choices. Consumers' erroneous inferences can be corrected by educating them about the shape of the distribution of liking differences. We discuss theoretical and managerial implications for the understanding and usage of polls.

Keywords: preference inference, liking inference, polls, consensus, social proof, preference cascades

In an autoguide.com (2017) poll in which 3000 consumers answered, “Which is better, the Accord or the Camry?”, 72% chose the Honda Accord over the Toyota Camry. One may infer from this poll that consumers like the Accord much more than the Camry. Surprisingly, this inference is not necessarily correct. The Accord may have received a considerably higher average rating than the Camry, but it is, in fact, much more likely that the two cars received very similar ratings.

In general, it is more likely that the difference in ratings of two alternatives is small rather than large. When ratings for each alternative are made on scales from 1 to 10, for example, it is much more likely to observe a difference of 1 than of 9 points because there are many more possible permutations of ratings resulting in a difference of 1 (e.g., 10 & 9, 9 & 8, 8 & 7, 7 & 6, etc.) as opposed to 9 (10 & 1). This permutational logic holds irrespective of what proportion prefers one car to the other or what scales (e.g., 5-point or 10-point) ratings are made on. Hence, even if 90% preferred the Accord to the Camry, it is much more likely that the two cars received very similar than vastly different ratings.

We demonstrate this regularity empirically using real-world data and with simulations. Our simulations show this is robust to the size and degree of the correlation between the options' liking of each group (e.g., those who prefer the Camry and those who prefer the Accord), and occurs for most distributions. Importantly, we show that consumers are not aware of this regularity, and believe that a large consensus implies large differences in liking. Consequently, they tend to overestimate how much better the majority preferred option is than the minority preferred option, which is one reason why polls and consensus information are so persuasive.

Our empirical findings suggest that managers, politicians, and public policy makers can elect how to aggregate ratings to shape consumers' judgments and choices. For instance, we conjecture that aggregating and displaying others' ratings as consensus (the percentage of people who prefer each option) rather than average ratings will make the majority preferred option seem more attractive, thereby increasing its choice share. Next, we elucidate our theoretical framework about how consumers infer differences in liking between the options from consensus information.

THEORETICAL DEVELOPMENT

Polls, Choices, and Differences in Liking

In the famous Pepsi Challenge of the 80s, the majority preferred Pepsi in blind taste tests over Coke. The Pepsi Challenge was very effective in changing consumer preferences and caused a major decline in Coke's market share (History.com 2020). In subsequent years, Coke came back with their own taste poll, and Pepsi is still doing taste challenges today, most recently showing that 71% of the Republic of Ireland prefers Pepsi Max over the market leader in caffeinated carbonated sodas (Pepsimax.ie 2019).

Polls are a popular method to gauge and convey aggregate beliefs in marketing (e.g., 71% prefer Pepsi Max, which restaurant is voted to have the best cheesesteak in Philadelphia; Phillymag.com 2018), sports (which NFL team is predicted to win the Superbowl; ESPN 2020), politics (75% of republicans approve of Donald Trump, Breitbart.com 2021), public policy (the majority of Americans support abortion rights; Forbes 2021), in short in any domain in which aggregate beliefs are of interest. Polls are popular for three reasons. First, they make it easy for respondents to express their beliefs since choosing is more natural and less cognitively demanding than evaluating options on rating scales (Fisher and Keil 2018; Fisher, Newman, and

Dhar 2018; Huber, Ariely, and Fischer 2002; Peterson and Pitz 1988). Second, poll outcomes are easily communicated with a single number (71% prefer Pepsi Max), whereas beliefs expressed as ratings, scores, or WTP require the comparison of two numbers, one for each option. Third, polls that result in a consensus allow marketers/policy makers to use a majority's opinion as a persuasion strategy. For example, "consensus messaging" such as informing the public that 97% of climate scientists agree that climate change is largely human-caused has been recommended as an effective communication strategy aimed at convincing people that climate change is real (Myers et al. 2015; Van der Linden et al. 2015).

While we focus primarily on polls about respondents' preferences, like the Coke-Pepsi example, our research extends to any domain in which a respondent's choice can be associated with an unobservable degree of belief about how much better (or truer) the majority chosen option/opinion/belief is. For ease of exposition, in this paper we will use the term *preference* to indicate which option a respondent would choose, the term *liking* (a continuous measure) to indicate how respondents feel about each option, and the terms *differences in liking* or *relative liking* to indicate how much a respondent likes one option compared to the other. When a sufficient percentage of respondents prefer one option over another, we will refer to the percentage preferring the majority preferred option as the "consensus" (e.g., 71% of people prefer Pepsi Max). We theorize about how consumers make inferences about the (unobservable) difference in liking between two options after learning the results of a poll and hence the consensus. In our empirical section, we demonstrate that small differences in liking are more likely than large differences—and test the resulting prediction that consumers will overestimate differences in liking inferred from consensus information—across a variety of domains such as liking of beers, wines, movies, and hotels, funniness ratings of jokes, predictions of sports games, and political party elections.

Our conceptualization is based on rational choice theory (Keeney, Raiffa, and Meyer 1993; Luce 1977) applied to forced choice, in which choices made in a poll are based on a comparison of the utility (or liking) that each choice option affords. Whichever option is liked more is chosen.[1] While differences in liking (an interval measurement) also tell us how the two options are ranked (an ordinal measurement), the opposite is not true. Choice or ranking of two options communicates little about the extent to which one option is liked more than the other. Consumers, however, may infer from consensus information the extent to which one option is liked more than the other. From “72% prefer the Accord over the Camry,” they may infer respondents liked the Accord much more than the Camry. Expressed in measurement terms, consumers infer from consensus information (72% prefer the Accord over the Camry)—an aggregate of ordinal data—the extent to which poll respondents like the Accord more than the Camry—a difference in aggregates of interval data (difference in average ratings).

Consensus Levels and Corresponding Differences in Liking

Consider a poll with two options, A and B, in which the liking of the options underlying a respondent’s choice is expressed in integers and is uniformly distributed between 1 and 10 (where 1 = lowest liking level and 10 = highest liking level; we will later relax these assumptions). Because polls usually require respondents to choose either A or B, we will only

[1] The assumption that choices reveal utilities, and that hence utility measurements (e.g., ratings of liking, WTP, or rankings) can be translated into choices, underlies many statistical models of choice (e.g., logistic regression, logit and probit models), commonly used tools in marketing (e.g., conjoint analysis, segmentation, cluster analysis), and most demonstrations of preference reversals. Furthermore, while there are many ways in which utilities might translate into choices, most polls are inherently forced choice. In forced choice tasks, utilities can only be translated to choice by selecting the option with the highest utility (like the mentioned examples do).

consider differences in liking that are not 0 (i.e., indifference by a respondent is precluded). Note that a respondent’s liking of A and B are not independent, having preferred one option (A) implies the other option (B) must be liked less. Table 1 below shows all possible differences in liking between A and B for a respondent that result from the 100 possible pairs. Any pair is as likely to occur as any other.

TABLE 1

POSSIBLE DIFFERENCES IN LIKING

	Liking of A									
Liking of B	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	1	2	3	4	5	6	7	8	9
2	−1	0	1	2	3	4	5	6	7	8
3	−2	−1	0	1	2	3	4	5	6	7
4	−3	−2	−1	0	1	2	3	4	5	6
5	−4	−3	−2	−1	0	1	2	3	4	5
6	−5	−4	−3	−2	−1	0	1	2	3	4
7	−6	−5	−4	−3	−2	−1	0	1	2	3
8	−7	−6	−5	−4	−3	−2	−1	0	1	2
9	−8	−7	−6	−5	−4	−3	−2	−1	0	1
10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0

NOTE.— All possible permutations of liking of two options A and B and their corresponding differences (liking values are uniformly distributed)

Differences in liking range from −9 (A = 1 and B = 10) to 9 (A = 10 and B = 1). There are nine ways for a respondent’s difference in liking to be 1 or −1, eight ways for a respondent’s difference to be 2 or −2, seven ways for a respondent’s difference to be 3 or −3, and so on. So, when looking at all possible cases in Table 1, small differences are more likely than large differences. This permutational logic holds not only for the individual respondent, but also aggregates up to the poll’s possible differences in liking across many respondents. Thus, for all polls and hence all consensus levels, whether 50% prefer A over B (all cases in Table 1) or 100%

prefer A over B (the 45 cases in the upper right grey area of Table 1) or 100% prefer B over A (the 45 cases in the lower left grey area of Table 1), small differences in liking are always more likely than larger differences.

From Table 1, we can also calculate the expected difference in liking for a given consensus level. To do so, we assume that instead of denoting differences of liking for a *single* respondent, the numbers in Table 1 denote the possibilities of a poll, that is the average difference in liking across *all* respondents in a poll. Let us take a look at the most extreme consensus level in which 100% prefer A over B (the 45 cases in the upper right grey area with positive differences). Because the underlying liking of the options is uniformly distributed, the expected difference in liking is simply the average of the 45 differences, that is 165/45 = 3.667. Thus, if 100% of respondents preferred A over B, the expected difference in liking would be 3.667. Conversely, if 100% preferred B over A, the expected difference in liking would be −3.667.

Building on this, we can calculate the expected difference in liking for different consensus levels by weighting liking differences by the relative size of the majority and the minority group (see Table 2). For example, when 90% prefer A over B and 10% prefer B over A, the expected difference in liking is 3.667 x 90% − 3.667 x 10% = 2.933.

TABLE 2

EXPECTED DIFFERENCES IN LIKING FOR DIFFERENT CONSENSUS LEVELS

Consensus	Expected difference for majority	Expected difference for minority	Overall expected difference in liking
100% A (0% B)	3.667 x 100%	−3.667 x 0%	3.667
90% A (10% B)	3.667 x 90%	−3.667 x 10%	2.933
80% A (20% B)	3.667 x 80%	−3.667 x 20%	2.201
70% A (30% B)	3.667 x 70%	−3.667 x 30%	1.467
60% A (40% B)	3.667 x 60%	−3.667 x 40%	0.733
50% A (50% B)	3.667 x 50%	−3.667 x 50%	0

NOTE.— All possible permutations of liking of two options A and B and their corresponding differences (liking values are uniformly distributed)

Table 2 shows that, as the level of consensus decreases, the expected difference in liking also decreases. In fact, the two correlate with r = 1. This corresponds well with intuition. The expected difference in liking for options A and B is likely to be larger when 90% of respondents prefer A over B than when only 60% do so. At *any* given consensus level, however, smaller differences in liking are more probable than larger differences due to the permutational logic outlined above. It is this key property that defies intuition. No matter whether 90% or 60% prefer A over B, it is more probable that both options are liked to a similar extent than it is that one option is liked vastly more than the other.

Are Small Differences in Liking Always more Likely than Large Differences?

Before we outline the consequences of this counterintuitive logic, let us briefly discuss the assumptions behind our modeling. In the above analyses, liking for the options assumed only integer values, were uniformly distributed, and positively correlated[2]. Relaxing the first assumption by allowing for decimals changes the distributions of expected liking differences only minimally (see the simulations in Web Appendix A). Relaxing the second assumption by assuming that moderate levels of liking are more likely than extreme levels (e.g., by switching from uniform to normal distributions truncated at the end points 1 and 10) causes smaller differences to become even more likely[3]. Relaxing the second and third assumptions by specifying the sign of the correlation between the liking of option A and B causes shifts in opposite directions. When the liking of option A and B within a group are positively correlated (e.g., the more respondents like A the more they also like B), smaller differences become more likely. When the liking of option A and B within groups are negatively correlated (the more respondents like A the less they like B), larger differences become more likely. But even in these cases, the effect of the permutational logic making small differences more likely than large differences still supersedes the effect of negative correlations. So, even when 90% prefer A over B and liking scores are correlated with −0.8, it is still the case that small differences in liking are more likely than large differences (see simulations in Web Appendix A).

What about when each group has a love-hate relationship with their preferred and non-preferred options (i.e., when liking differences follow bimodal distributions)? In these cases, the

[2] Within groups, that is for those who prefer A over B and those who prefer B over A, average liking for A and B correlate with 0.5 in our uniform simulations due to preference constraints. The intuition is as follows: imagine we graph liking such that the x-axis displays liking of A and the y-axis liking of B. For those who prefer A over B, the graph will show a left-skewed distribution (because liking of A must be greater than liking of B). This creates a correlation of liking between A and B of 0.5. Now imagine the same graph for those who prefer B over A. This graph will show a right-skewed distribution (because preferences for B must be greater than for A), resulting again in a correlation of 0.5 of liking scores. The correlation between A and B for each group can change when only a subset of all possible permutations is represented.

[3] The curious reader might also wonder what happens when distributions are J-shaped, such as in online reviews in which self-selection is present (Schoenmueller, Netzer, and Stahl 2020). We also simulated this case (see R code) and find the same general pattern of results. Additionally, our secondary datasets listed in the next section contain online reviews with the exception of the jokes data.

permutational mechanism breaks down because the liking of the preferred option is forced to be far from the liking of the non-preferred option for each poll respondent, and thus so too are the mean likings. For demonstration purposes, let's consider an extreme case in which liking of A and B are maximally different and highly negatively correlated. Imagine the 90% majority prefer A over B, with mean liking levels for A = 9/10 and for B = 1/10, while the 10% minority prefer B over A with opposite liking levels. We simulated this case with liking of A and B within each group correlated at -.82 (see details in Web Appendix A), resulting in large differences in liking being more likely than small differences.

Empirical Demonstration: Small Differences in Liking are More Likely than Large Differences

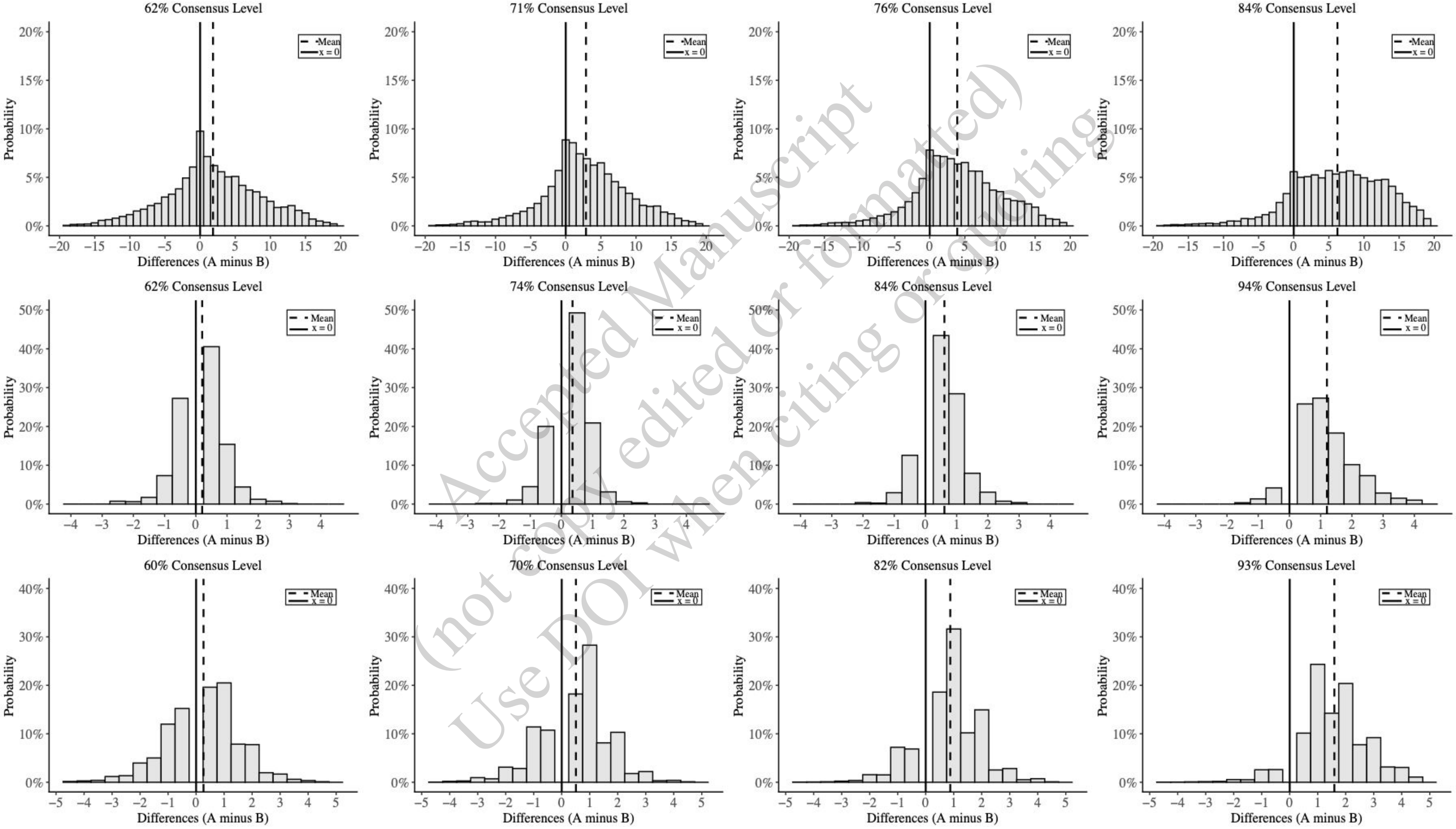
To test our logically derived hypothesis empirically, we acquired three datasets (jokes, beers, and movies) in which respondents rated objects. The jokes dataset contains more than 600,000 ratings of 100 jokes from 24,938 respondents, in which the median respondent rated 24 jokes on a scale from −10 to 10 (Goldberg, Roeder, Gupta, and Perkins 2001). The beer dataset contains more than 1.5 million ratings of 56,000 beers from 33,388 respondents, in which the median respondent rated 3 beers on a scale from 1 to 5 in .5 increments (https://www.kaggle.com/rdoume/beerreviews). The movie dataset contains more than 25 million ratings of 62,423 movies from 162,541 respondents, in which the median respondent rated 71 movies on a scale from 0.5 to 5 in .5 increments (Harper and Konstan 2015; https://grouplens.org/datasets/movielens/25m/). For more details about the datasets please see the Web Appendix B.

We selected ratings of target objects (liking and funniness) only from respondents who had rated both objects. This allowed us to calculate average differences in ratings (i.e., relative liking or relative funniness) as well as consensus levels, the proportion of respondents preferring one option over the other according to their ratings. Like in our analyses above, we eliminated ratings resulting in a zero difference to mimic choices in polls in which respondents are forced to choose one option over another. We obtain supporting evidence for the two key insights found in our analysis above and in our simulations (see Figure 1). First, we observe that small differences in ratings are generally more likely than large differences [4]. Second, a comparison of average differences (vertical dashed lines) across graphs within a row reveals that as consensus increases, so do average differences in ratings (i.e., they are highly positively correlated). For example, we see that for the joke pair producing a 62% consensus level (row 1, leftmost graph) the average difference in funniness is 1.8 while for the pair producing the 84% consensus level (row 1, rightmost graph) the average difference is 6.2.

[4] The rating differences for movies are particular in that respondents tend to rate movies in integers but not fractions, which explains why integer differences are more likely than fractional differences in movie ratings (cf., analysis in the Web Appendix B).

FIGURE 1

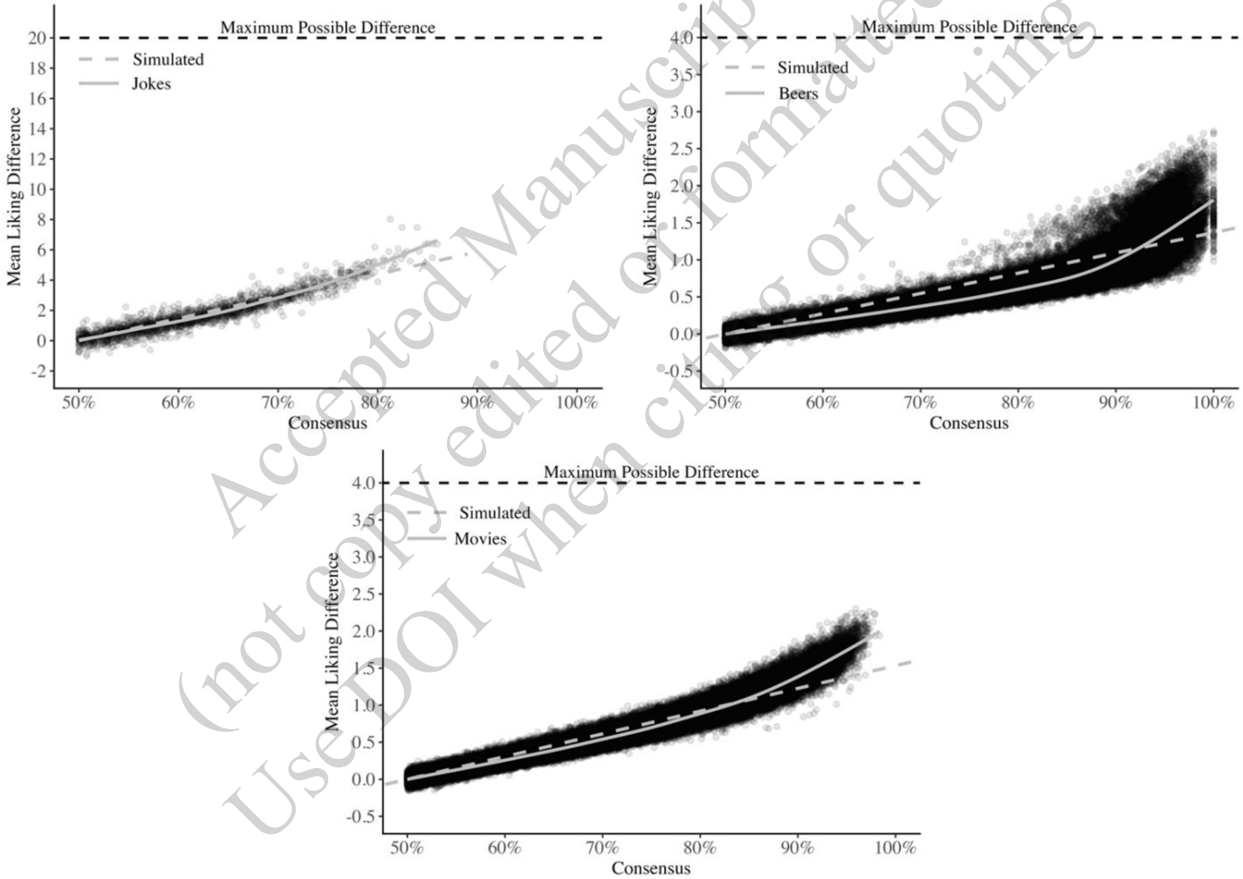
HISTOGRAMS OF RATING DIFFERENCES FOR JOKES (TOP), BEERS (MIDDLE), AND MOVIES (BOTTOM)



NOTE.— Figure 1 shows the distribution of rating differences for four different joke, beer, and movie pairs each, representing low to high consensus levels.

To further test the above observations, we examined thousands of possible joke combinations and more than 150,000 combinations of beers and movies. For each combination, we calculated the consensus, mean, and modal value (see Web Appendix C). We then compared the mean and mode estimates at different consensus levels to those derived from our simulations. As can be seen in Figure 2, the mean differences in ratings are small and extremely close to those derived in our simulations. The same holds for observed and simulated modal values.

FIGURE 2
SIMULATED AND ACTUAL MEAN DIFFERENCES IN RATINGS



NOTE.— Scatterplots of mean differences in liking for pairs of jokes, beers, and movies, on their respective scales. Darker areas indicate greater density of pairs with the same mean difference. The grey dashed lines display simulated mean differences, and the grey solid lines are the (actual) observed mean differences. The black dashed line marks the maximum possible difference that could occur on the scale for that dataset.

Given Consensus Information, Consumers Overestimate Average Differences in Liking

We surmise that when given consensus information such as 90% prefer A over B, consumers tend to infer from such information a corresponding difference in liking between A and B, that is by how much is A better than B. To do this, they follow a process similar to probabilistic consistency (Dick, Chakravarti, and Biehal 1990; Broniarczyk and Alba 1994), inferential imputation (Jaccard and Wood 1988; Kardes, Posavac, and Cronley 2004; Johnson and Levin 1985) or interattribute inference (Evangelidis and van Osselaer 2018, 2019). In all three processes, consumers infer the most likely value on an unknown dimension from the magnitude of the value on a known dimension and the perceived relationship between the two. For example, when consumers do not know a brand's quality but know its price, they may rely on their perception about the relationship of price and quality—i.e., quality and price are strongly positively correlated— and that high prices often have high quality to infer the brand's quality (Dick, Chakravarti, and Biehal 1990). In our case, the information to be inferred does not pertain to specific attributes of the stimuli but rather to metrics of liking (or funniness or superiority). We surmise that consumers infer poll respondents' (unobservable) relative liking (e.g., how much more do they like A than B?) from the magnitude of the reference information that is directly observable (e.g., consensus information) and the perceived correlation between the target and the reference. And since the correlation between consensus levels and average differences in liking is close to 1, consumers tend to perceive the two forms of liking/preference information to be substitutes (Kahneman 2003; Kahneman and Frederick 2002; Morewedge and Kahneman, 2010). From 50% consensus, they are likely to infer an average difference in liking close to zero, and as consensus levels increase, so, too, will the inferred average differences.

However, because consumers are unaware of the permutational logic according to which small differences in liking are—at any consensus level—more likely than large difference, they will infer distributions of liking differences that deviate from the truth, such that small differences are underrepresented and large differences are overrepresented. As a consequence, when given consensus information, consumers will tend to overestimate the mean and mode of differences in liking, believing the majority preferred option to be much better than it actually is compared to the alternative.

Effects of Larger Versus Smaller Magnitudes

We are not the first to examine consumers' beliefs about how the magnitude of a piece of information is viewed. Kupor and Laurin (2020) show that products whose outcomes have a larger probability of occurring are perceived to have higher rates of that outcome. For example, participants who learned that 68% of Claritin users experience coughing thought that someone taking Claritin would cough more than if they learned that 7% of users experience coughing. Westwood et al. (2019) show that voters who learn the probability that the leading candidate will win is 87% predict the leading candidate will have a larger share of the votes than when voters learn that the corresponding expected vote share for the leading candidate is 55%. In both papers, consumers viewing larger magnitudes (i.e., probability of occurrence) associated them with larger outcomes. Similarly, the literature on anchoring also predicts that people who are exposed to larger magnitudes will give larger estimates (Tversky and Kahneman 1974). In contrast to the above research which compares two different states, we make predictions about how a *single* magnitude value (consensus) can lead one to overestimate another value (difference in liking). Specifically, unlike the aforementioned research, our core prediction is not that large values of consensus lead to larger inferred differences in liking (which is true in expectation). Rather, our core prediction is that, when provided with information about a given consensus level, consumers will overestimate the unobserved difference in liking (compared to the true difference in liking). To this point, our theory specifies a process that does not require a comparison of two states (e.g., differences in inferences of liking between two consensus levels). We can however use our theory to also make additional predictions about comparative states.

Empirical Overview (Experiments 1-7)

We provide evidence for our hypotheses in seven experiments. Experiments 1 and 5 use our simulations as benchmarks, while Experiments 2, 3, 4, 6, and 7 do so with real-world data. Further, Experiments 1, 2, 3, and 4 focus on the domain of preferences, while Experiment 6 extends our findings to sports game predictions, and Experiment 7 to the prediction of political elections (for an overview see Table 3).

All experiments were preregistered. Correct answers were monetarily incentivized in Experiments 1, 3, 4, 6, and 7. All stimuli, experimental materials, code, data, and preregistrations are accessible at https://researchbox.org/446.

TABLE 3

EMPIRICAL OVERVIEW (EXPERIMENTS 1-7)

	Sample	Stimuli	Benchmark	IV	DV	Result
Experiment 1	86 students in a data analytics course at a European University	500 people rated two Wines A and B on scales from 1 (poor quality) to 6 (high quality)	Simulations	2 consensus conditions (btw-sbj) *60%* or *90%* rated wine A higher than wine B	Most likely average difference in liking of wine A and wine B	75% in *60% consensus* and 93% in *90% consensus* overestimated the most likely average difference in liking
Experiment 2	402 Prolific workers	501 Prolific workers rated funniness of two jokes	Funniness ratings of 501 Prolific workers	2 conditions (btw-sbj) *Control* (61% rated Joke A as funnier than Joke B) vs. *Learn and rate* (first read and rated the jokes, then estimated average difference in funniness from 61% consensus information)	Most likely average difference in funniness of Joke A and Joke B	81% in C*ontrol* and 78% in *Learn and rate* condition overestimated most likely average difference in funniness
Experiment 3	200 Prolific workers	601 beer ratings for Pliny the Elder and 120 Minute IPA, and 638 beer ratings for Old Rasputin and Fat Tire Amber	Dataset of 1.5 million beer ratings from Beeradvocate.com	2 consensus conditions (within-sbj) *74%* favored Old Rasputin over Fat Tire Amber and *94%* favored Pliny the Elder over 120 Minute IPA	Distribution builder for differences in liking	Overestimation of • Average difference in liking • Mode of difference in liking • Probability of maximum difference in liking Underestimation of • Probability of the mode of difference in liking • Probability of smallest positive difference in liking

Experiment 4	300 Prolific workers	601 beer ratings for Pliny the Elder and 120 Minute IPA	Dataset of 1.5 million beer ratings from Beeradvocate.com	2 conditions (btw-sbj) *Control* (94% favored Pliny the Elder over 120 Minute IPA) vs. *Debiasing* (first see distribution of liking differences for beer pair with 50% consensus, then build distribution for 94% favored Pliny the Elder over 120 Minute IPA)	Distribution builder for differences in liking	*Control:* results from Experiment 3 are replicated. *Debiasing*: over- and underestimations are greatly reduced or eliminated
Experiment 5	450 Amazon Mechanical Turk workers	Hotel A costs $125 a night, Hotel B costs $98 a night	Simulations	3 conditions (btw-sbj) *Average ratings condition* (Hotel A rated 87 out of 100 and Hotel B 75 out of 100) vs. *Percent consensus condition* (70% of people rated Hotel A higher) vs. *Out of consensus condition* (7 out of 10 rated Hotel A higher)	Choice between hotel A and hotel B	Choice shares of Hotel A: *Average ratings condition:* 42% *Percent consensus condition:* 76% *Out of consensus condition:* 79%
Experiment 6	400 Amazon Mechanical Turk workers	Super Bowl games LI and LIII	Actual point spreads as provided by ESPN experts	2 consensus conditions (btw-sbj) *72%* (Super Bowl LI) or *62%* (the Super Bowl LIII)	Estimation of average point spread by ESPN experts	Overprediction of point spreads: *62% consensus:* $M_{\text{predicted}}$ = 13.1, M_{actual} = 1.0 *72% consensus:* $M_{\text{predicted}}$ = 13.6, M_{actual} = 1.9
Experiment 7	600 Prolific workers	Election in a European country	Actual vote share (54.1%) that the dominant party received in the election	2 conditions (btw-sbj) Participants were provided with information from a nationally representative poll of 1269 citizens: C*onsensus information* vs. *Liking ratings* for the 10 parties	Prediction of vote share of the dominant party	Lower vote share predictions from *Liking ratings* M_{ratings} = 38.3 than from C*onsensus information* $M_{\text{consensus}}$ = 43.0

EXPERIMENT 1: OVERESTIMATION OF AVERAGE DIFFERENCES WITH SIMULATED DATA

Experiment 1 tested our hypothesis that participants overestimate differences in liking when presented with consensus information. Participants were asked to predict the most likely difference in average wine ratings for two wines, A and B, in which either 60% or 90% of wine tasters had given Wine A a higher rating than Wine B. To make Experiment 1 a conservative test of our hypotheses, we recruited a group of analytically trained students from a master's program at a European business school who were taking a data analytics course, and incentivized accuracy by offering 25 euros each to two students who gave the correct answer.

Method

We aimed at recruiting 100 participants and ended up with 86 (67.4% female, M_{age} = 23.3, *SD* = 1.7) from the data analytics class on the one day the study was run. Participants were randomly assigned to either a low consensus (60%) or a high consensus (90%) condition. Specifically, participants were informed that "500 people rated both Chardonnays below (Wine A and Wine B) on a scale from 1 to 6 (1 = poor quality and 6 = very high quality)." In the low [high] consensus condition participants learned that "60% [90%] rated Wine A as higher quality than Wine B. 40% [10%] rated Wine B as higher quality than Wine A." A comprehension check ("What rating scale was each wine rated on?") had to be answered correctly to advance with the study. Participants were then asked which difference between the ratings of Wine A and Wine B they thought was the most likely. The answer options were:

a) The average rating of Wine A is about 0-1 point higher than that of Wine B,

b) The average rating of Wine A is about 1-2 points higher than that of Wine B,

c) The average rating of Wine A is about 2-3 points higher than that of Wine B,

d) The average rating of Wine A is about 3-4 points higher than that of Wine B,

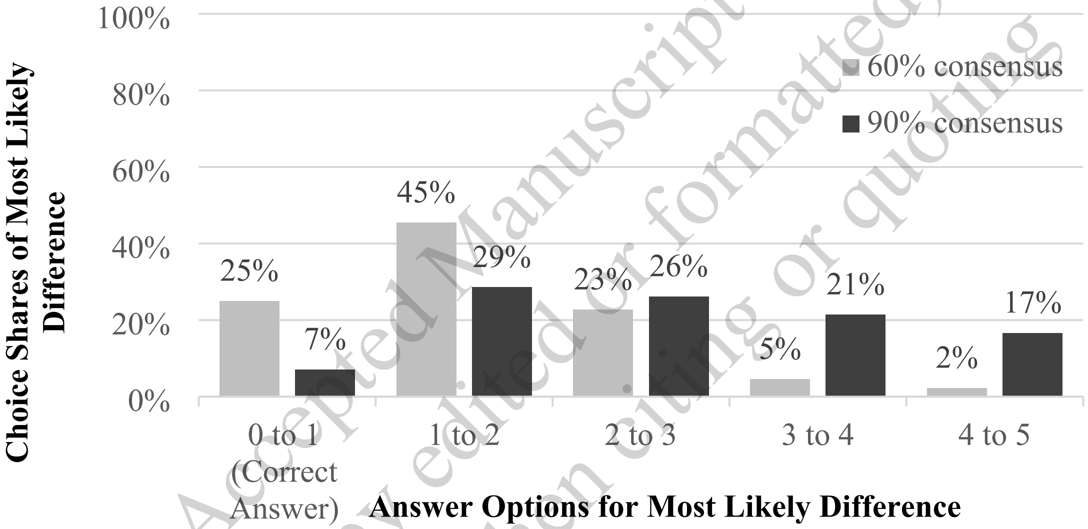
e) The average rating of Wine A is about 4-5 points higher than that of Wine B."

Note that each choice option comprises a range of values rather than asking about a specific value. We did this because we thought it would be easier to compare likelihoods of ranges of values than likelihoods of specific values. We predicted that: 1) in both the 60% and 90% consensus conditions, a minority of participants would choose the correct answer option (a) as observed in our simulations, 2) in the 90% consensus condition compared to the 60% consensus condition, more participants would choose larger difference answer options, (e) and (d), than relatively smaller difference answer options, (b) and (c). Participants were informed that there was only one correct answer. If they guessed it, they would be entered into a draw in which two winners would each win 25 euros.

Results and Discussion

Results are summarized in Figure 3. Participants in both conditions believed that larger differences in ratings were more likely than smaller ones. Importantly, as predicted, in the 60% consensus condition, the majority of participants (75%) believed a larger difference was more likely than the incentivized, correct answer of "about 0-1" (25%; Pearson $\chi^2(1) = 11.00$, $p = .001$). Likewise, in the 90% consensus condition, as predicted, the majority of participants (92.9%) chose larger differences to be more likely than the incentivized, correct answer of "about 0-1" (7.1%, Pearson $\chi^2(1) = 30.86$, $p < .001$). As a conservative test, we also tested whether the majority of participants in the 90% consensus condition chose an answer greater than "about 1-2" compared to choosing the lowest two ranges of "about 0-1" and "about 1-2", they did (64%, Pearson $\chi^2(1) = 3.43$, $p = .064$). Finally, as predicted more participants in the 90% than the 60% consensus condition believed that the largest differences of "about 3-4" and "about 4-5" were more likely than the smaller differences of "about 1-2" and "about 2-3" (41% vs. 9.1%, Pearson $\chi^2(1) = 9.38$, $p = .002$).

FIGURE 3
RESULTS OF EXPERIMENT 1



NOTE.— Proportion of respondents choosing each answer option for the most likely difference in ratings in Experiment 1.

Participants in Experiment 1 overestimated the modal difference in ratings as they predominantly failed to choose the correct answer of a 0 to 1 rating point difference despite being incentivized to do so. This occurred even with master level students that had taken several classes in data analytics.

EXPERIMENT 2: INFERRING RATINGS AND THE ROLE OF PRIORS

Experiment 2 is a conceptual replication of Experiment 1 with joke ratings. Instead of using our simulations as a normative benchmark, we collected our own benchmark data. In a first stage, participants read two jokes taken from our jokes dataset, chose which of the two jokes they found funnier, and then rated each joke's funniness. In a second stage, a different set of participants was provided with the consensus information calculated from the first stage and asked to estimate how much funnier they thought one joke was rated than the other.

Method

Stage One. We recruited 501 participants via Prolific to complete stage 1 of our study online in exchange for £0.30. Following our preregistration, we dropped anyone who took our survey more than once or gave an answer beyond the range restricted by our Qualtrics coding, leaving us with 499 participants (47.5% female, M_{age} = 29.7, SD = 9.6). To select two jokes from the joke dataset, we looked at pairs in which each joke had more than 10,000 ratings and eliminated jokes that might be offensive (e.g., sexual jokes and black humor). We finally selected a joke pair that had been rated by 15,096 respondents. Of the 15,096 respondents, 71% had rated a joke about a group of managers trying to measure the height of a flagpole as funnier than another joke about a dog sending a telegram (the jokes can be found in the survey files on Researchbox). Participants first read each joke (in random order), chose which joke they thought was funnier, and then rated how funny they thought each joke was on a scale from 1 "Not Funny" to 10 "Very Funny". Finally, they reported demographics.

Stage One Results. We used the choice data to determine the consensus level. 61% percent chose the joke about the managers as funnier, while 39% chose the joke about the dog as funnier. We then took the ratings of the dog joke, henceforth Joke B, and subtracted them from the ratings of the manager joke, henceforth Joke A. The mode of the resulting distribution of differences, Joke A minus Joke B, served as the normative benchmark for the estimated difference in ratings by stage two participants (see below). We observed that only 14 of the 499 participants gave the same rating to both jokes and thus expressed indifference. No participants rated their less-preferred joke higher than their more-preferred joke (i.e., no preference reversals between choices and ratings were observed). Excluding the ratings/choice of the 14 indifferent participants does not, in any meaningful way, affect the consensus level or modal difference. Furthermore, calculating consensus from ratings produces virtually identical choice shares to those derived directly from choice: 59% rated Joke A as funnier, 38% rated Joke B as funnier, and 3% rated them equally.

Stage Two. We recruited 402 participants (who did not participate in stage 1) via Prolific to complete our study online in exchange for £0.30 (52.2% female, Mage = 29.2, SD = 9.16). Participants first read that "We are interested in how funny you think two jokes, Joke A and Joke B, are. Both jokes were shown to 499 participants in a previous study. The participants indicated which joke they thought was funnier, and also rated how funny each joke was on a scale from 1 to 10, "Not Funny" to "Very Funny"." Participants were then randomly assigned to one of two conditions. In the *control condition*, participants learned that 61% chose Joke A as funnier, 39% chose Joke B as funnier. As in the first stage, the labels and order of presentation of the jokes were counterbalanced. In the *learn and rate condition*, participants first read both jokes (presented in random order and with a counterbalanced label) and after reading each joke, rated how funny they found each joke to be using the same scale as stage one participants. Then they were given the same consensus information as in the control condition. Participants in both control and learn and rate conditions were then asked, "What do you think is the most likely difference in Joke A's rating minus Joke B's rating produced by the participants in the previous study?" Participants could enter a value between −9 and 9, and were given two examples, one illustrating Joke B being funnier than Joke A resulting in −9 and the other illustrating Joke A being funnier by 9 points. Finally, they answered demographic questions and completed the survey.

Comparing the *learn and rate condition* to the *control condition* allowed us to test whether forming prior beliefs would influence estimates of the average difference in funniness ratings. It also ensured that participants underwent a similar task to that of stage one participants, which presumably made it easier for them to estimate stage one participants' liking differences.

We predicted that stage 2 participants in the *control condition* would overestimate differences in ratings in stage 1, specifically, that the majority of participants would overestimate the median of 1 observed in stage 1 ratings, and that the average estimate of the most likely difference in ratings would be greater than the median. We had expected the median to be higher than the mode, which is why we had intended to use the median as a more conservative benchmark. It turned out we were wrong, the median in stage 1 ratings was actually lower than the mode of 2. In our below analyses, we hence deviate from our preregistration and use the more conservative benchmark of mode = 2 for testing overestimation. We did not make formal predictions about the *learn and rate condition*.

Stage Two Results and Discussion

As predicted, the majority of participants in the *control condition* overestimated the most likely difference in ratings in stage 1, entering a value higher than the observed mode of 2 (81.2%, Pearson $\chi^2(1) = 78.59$, $p < .001$). So too did those in the *learn and rate condition*

(77.5%, Pearson $\chi^2(1) = 60.50$, $p < .001$). The percent overestimating did not differ significantly between the two conditions (81.2% vs. 77.5%, Pearson $\chi^2(1) = 0.62$, $p = .429$).

As predicted, the average modal estimate in the *control condition* was significantly higher than the actual mode of 2 ($M_{\text{control}} = 4.24$, $SD = 2.78$, $t(201) = 11.47$, $p < .001$). The same was true in the *learn and rate condition* ($M_{\text{learn\_and\_rate}} = 3.80$, $SD = 3.41$, $t(199) = 7.44$, $p < .001$). Though directionally lower, the *learn and rate condition* did not differ from the *control condition* in the degree of overestimation ($M_{\text{control}} = 4.24$, $SD = 2.78$ vs. $M_{\text{learn\_and\_rate}} = 3.80$, $SD = 3.41$, $t(400) = 1.44$, $p = .150$).

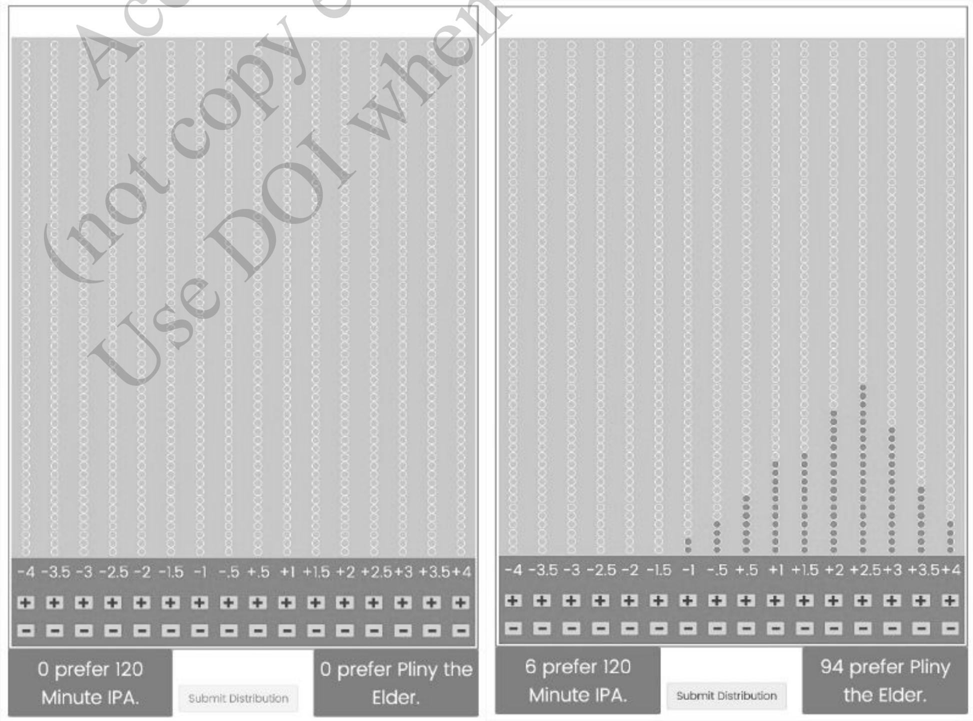
The results corroborate the findings of Experiment 1 that participants overestimate the difference in liking when inferring it from consensus information. Participants in Experiment 2 overestimated the difference in funniness ratings for two jokes when learning that 61% preferred Joke A over Joke B.

One limitation of Experiment 2 is that we sampled only one pair of jokes from the 100 jokes that had been rated by 24,938 respondents in our joke dataset. The hypothesized and observed overestimation may hence occur only for a few pair of jokes but not generalize across all joke pairs. We address sensitivity to stimulus selection in Experiment 3 and in Appendix E.

EXPERIMENT 3: ELICITING DISTRIBUTIONAL BELIEFS ABOUT DIFFERENCES IN LIKING

In Experiment 3, we examined the source of consumers' overestimation by eliciting their beliefs about how differences in liking are distributed. To this end, we showed participants the results of two polls from our real-world beer data and asked them to construct the entire distribution of liking differences for poll responders. By doing so, we could track for each participant whether they over- (or under-) estimated mean differences in liking of the options. Distributions of liking differences were elicited with the distribution builder, see Figure 4 (Sharpe, Goldstein, and Blythe 2000; Goldstein and Rothschild 2014; quentinandre.net 2022). As before, we predicted that participants would overestimate how much better option A is than option B. Specifically, we expected participants to shift distributions too much to the right, thereby not only overestimating the mean, mode, and probability of the maximum difference in liking, but also underestimating the probability of the mode and of small differences in liking. Experiment 3 also allowed us to test the sensitivity of our results with respect to the specific stimuli that we had chosen. To this end, we conducted a counterfactual sensitivity analysis testing for how many of all possible beer pairs with similar consensus levels we would have observed overestimation by participants in Experiment 3.

FIGURE 4
EXAMPLE OF DISTRIBUTION BUILDER TASK IN EXPERIMENT 3



NOTE.— Examples of the distribution builder tool used in Experiments 3 and 4. The left panel is participants' starting point and the right panel represents a participant's final allocation.

Method

We recruited 200 participants via Prolific to complete our online study in exchange for £0.90 (48% female, M_{age} = 27.63, *SD* = 9.67). All participants were shown a beer pair with one of two consensus levels, 74% and 94%, one at a time, and were randomly assigned to the order in which they viewed the two beer pairs.

Our stimuli were derived from the beer dataset mentioned in our introduction. To select ratings of four beers to form two beer-pairs with different consensus levels, we looked at all beer-pairs in which each beer had more than 500 ratings (in order to derive more precise estimates of rating distributions). From that list of beer-pairs, we chose Pliny the Elder and 120 Minute IPA for which 601 users rated both resulting in a 94% consensus favoring Pliny the Elder, and Old Rasputin and Fat Tire Amber for which 638 users rated both resulting in a 74% consensus level favoring Old Rasputin.

Participants saw either the beer pair with 94% or with 74% consensus level first. For each consensus level beer-pair, participants read that "100 users on www.beeradvocate.com rated the two beers shown below on a scale from 1 to 5 with 0.5 increments (1 = worse rating and 5 = best rating)," and were provided with images of the two beer bottles. For example, when viewing the 94% consensus level beer pair, participants learned that "94% gave Pliny the Elder a higher rating than the 120 Minute IPA" and "6% gave the 120 Minute IPA a higher rating than Pliny the Elder." Below this they were asked: "What do you think the distribution of ratings looks like in this case?" Participants were then instructed to use the distribution builder to allocate each of the 100 users to a difference in ratings, which in the 94% case read "Pliny the Elder's rating minus 120 Minute IPA's rating." They were informed that a positive [negative] rating means that a user likes Pliny the Elder more than 120 Minute IPA [or vice versa] and the greater this positive [negative] value, the greater is the difference in liking. Participants learned they could add and subtract users with the buttons below the x-axis, and that their allocation must match the consensus level in order for them to proceed. Participants were finally told that if their distribution matched the actual distribution, they would receive a 20-cent bonus. After building preference distributions for both consensus level beer-pairs, participants provided their demographic information.

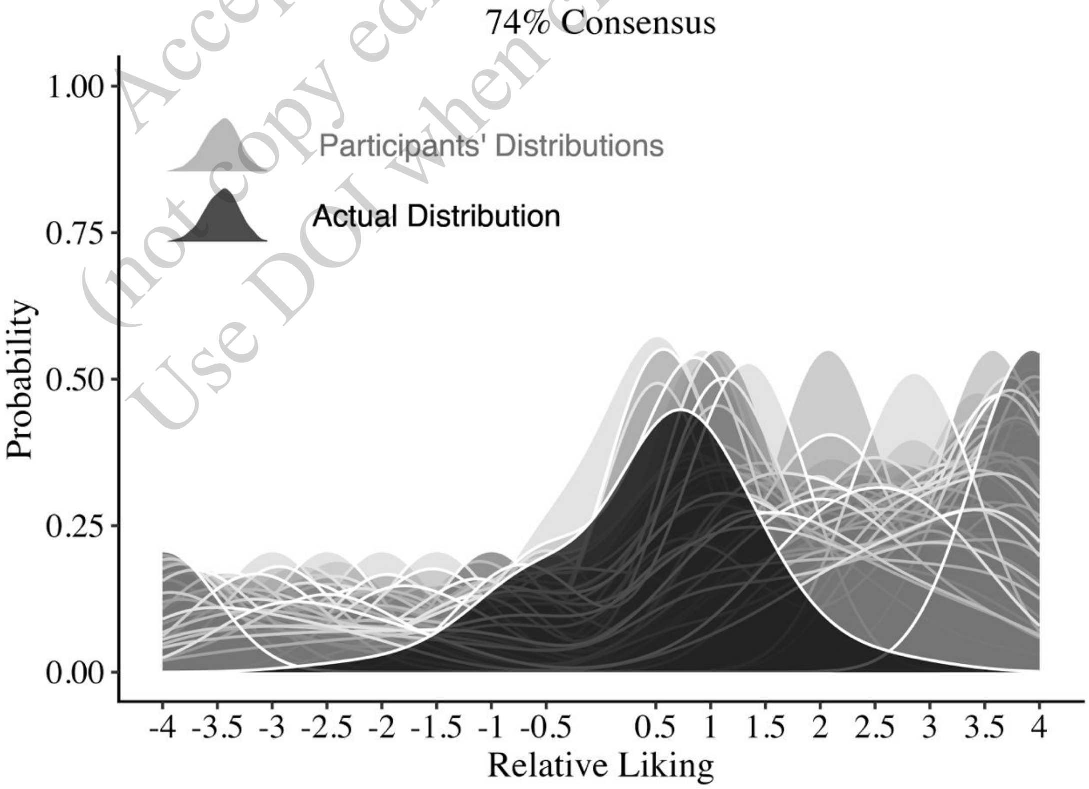
To make the task less demanding, we asked participants to allocate 100 users from each consensus level rather than the 601 and 638 raters upon which the 94% and 74% consensus level ratings were based, respectively. For each observed participant distribution we computed the mean, mode, probability of the mode, probability of the smallest positive difference, and probability of the maximum difference in liking.

When building distributions, one consequence of asking participants to allocate 100 users instead of the total number of users from the actual beer data, is that a participant cannot easily replicate the true, full distribution for a given consensus level. As a solution, we created benchmarks as follows: we randomly drew 100 users (rating differences) from the true distributions subject to the corresponding consensus level constraint, and repeated this 10,000 times for each consensus level. This resulted in two datasets, one for each consensus level. From these datasets, we computed the parameters of interest with their corresponding 95% confidence intervals. We classified a participant's parameter estimates as an overestimation when they exceeded the 95th percentile, and as an underestimation when they fell below the 5th percentile.

Results and Discussion

A summary of the results is provided in Table 4, for the full results see Appendix D Table S1. Figure 5 graphs every participant's estimated distribution (in grey) and the actual distribution (in black), for both the 74% and 94% consensus levels. Like in our previous studies, participants overestimated mean differences in liking for both consensus levels. Furthermore, as predicted, participants overestimated the mode and the probability of the maximum difference, and underestimated the mode's probability and the probability of the smallest positive difference (i.e., 0.5). A majority of participants committed each of these errors. No order effects were observed for any of the estimates. These results can be seen visually in Figure 5 as the mass of most participants' distribution is to the right of the actual distribution, along with the peaks (i.e., mode). Relatedly, not enough mass was given to small values (like the mode and minimum positive value).

FIGURE 5
ESTIMATED DISTRIBUTION VS ACTUAL DISTRIBUTION IN EXPERIMENT 3



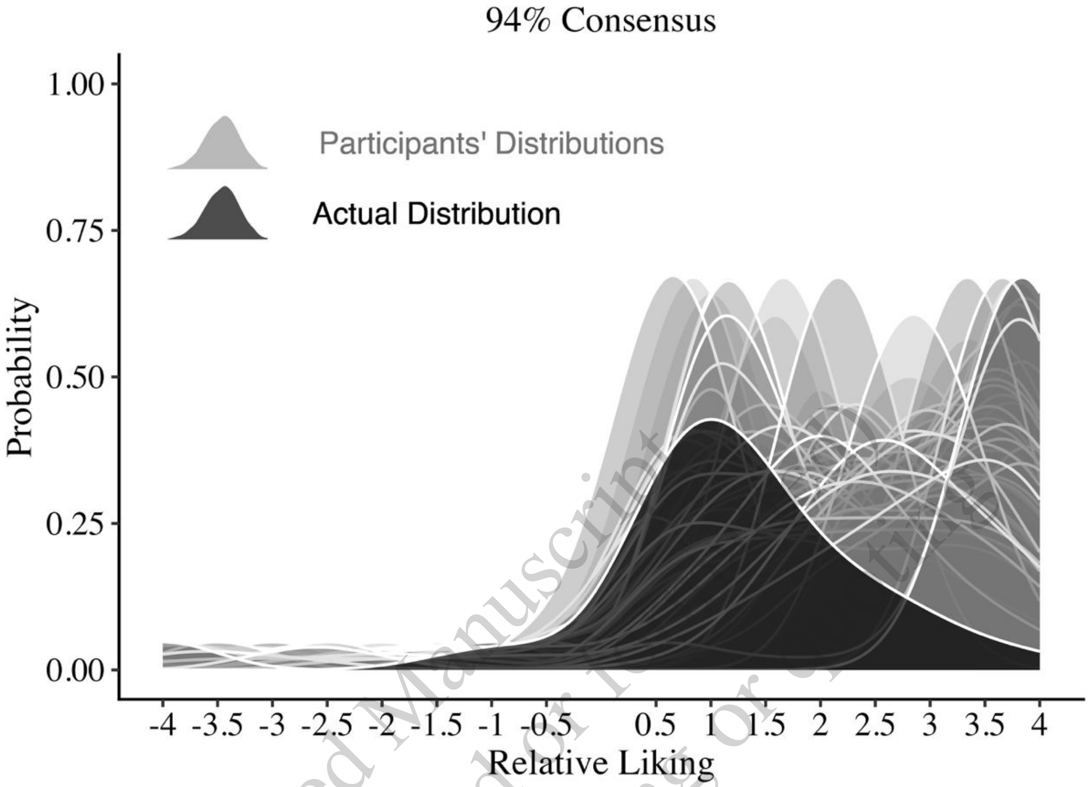


TABLE 4

SUMMARY OF THE RESULTS OF EXPERIMENTS 3 AND 4

		Experiment 3		Experiment 4	
	DV	74% Consensus	94% Consensus	94% Consensus Control	94% Consensus Debias
over-estimation	mean	90.5% *p* < .001	94.0% *p* < .001	94.7% *p* < .001	60.7% *p* = .011
	mode	75.5% *p* < .001	85.5% *p* < .001	77.3% *p* < .001	40.0% *p* = .165
	probability of maximum difference	82.0% *p* < .001	81.5% *p* < .001	78.0% *p* < .001	43.3% *p* = .121
under-estimation	probability of mode	98.0% *p* < .001	90.0% *p* < .001	88.7% *p* < .001	59.3% *p* = .027
	probability of smallest positive difference	98.0% *p* < .001	93.0% *p* < .001	90.0% *p* < .001	51.3% *p* = .807

Note—Proportion of participants who over- and underestimated distribution parameters together with chi-square tests against 50%.

Counterfactual Sensitivity Analysis regarding Stimulus Sampling

One may wonder how sensitive our results are to the choice of stimuli. It could be that the beer pairs we used in this study have particularly low means and modes, making it much more likely we would find our results, and if we instead had used a different pair of beers our results would not be similar. We address this concern by calculating the normative benchmarks for all beer pairs we could have chosen from our dataset that would have produced a consensus level similar to 74% or 94%, like in our study. For the 4,006 beer pairs producing a consensus level of 74%, 99.9% had a lower mean and mode than what participants had predicted in Experiment 3. For the 2,319 beer pairs producing a consensus level of 94%, 96.6% had a lower mean and 82.6% had a mode lower than what participants had predicted in Experiment 3. Note that by using the estimates from Experiment 3 as our benchmark, we assume that participants' answers are driven solely by consensus levels and are independent of the specific beer brands, labels, and bottles. For more details about the counterfactual sensitivity analysis, please see Appendix E. This analysis shows that the overestimation of the mean and the mode of average rating differences are robust across almost all possible pairs with the given consensus level.

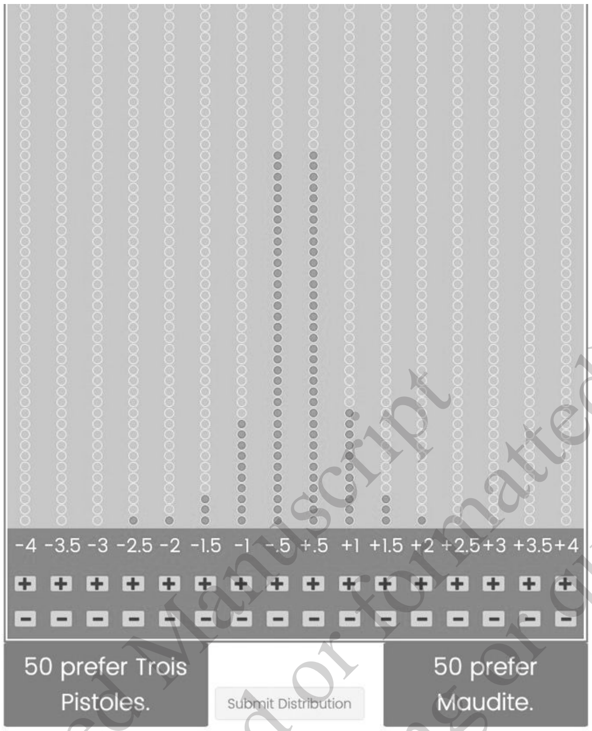
Using another real-world dataset, Experiment 3 provides further evidence that participants infer from consensus information differences in liking that are too large. Rather than asking participants directly to estimate mean differences, in Experiment 3, participants' beliefs were revealed by the distributions of differences that they built. The elicited distributions show that it is the underestimation of the likelihood of small differences that causes participants not only to overestimate mean differences, but also the likelihood of maximum differences.

Note that we do not believe that people encountering polls spontaneously conjure a distribution and draw their inferences about preference differences from it. Instead, we use the distribution builder to capture the visceral feeling when seeing that 94% prefer Pliny the Elder: hot damn, that must be a much tastier brew! Importantly, it should be noted that distributions for the 74% and 94% consensus levels did not depend on which distribution participants were asked to build first. The absence of an order effect suggests that participants are not naively anchoring on the consensus level they had seen before, but instead seem to independently infer difference in liking when being confronted with consensus information.

EXPERIMENT 4: DEBIASING RESPONDENTS

In Experiment 4, we test whether showing participants the preference distribution for a 50% consensus case will make them realize that, in general, small differences in liking are more likely than large differences, and thus help reduce overestimation of mean differences for consensus levels other than 50%. As in Experiment 3, we asked participants to build the distribution of liking differences for the 94% consensus beer-pairing. Participants in the *debias condition* were first shown the distribution of liking differences for the beer pair Trois Pistoles and Maudite, which displayed a 50% consensus level (participants in the *control condition* were not shown this distribution). This 50% consensus distribution was nearly symmetrical with descending staircases away from zero: one into the positive numbers for those who prefer beer A and one into the negative numbers for those who prefer beer B (see Figure 6).

FIGURE 6
THE 50% CONSENSUS DISTRIBUTION DISPLAYED FOR HALF OF THE PARTICIPANTS IN EXPERIMENT 4



50 prefer Trois Pistoles.

50 prefer Maudite.

There are three likely ways in which participants in the *debias condition* might use the 50% consensus distribution of liking differences to draw inferences about the 90% consensus distribution produced by Pliny the Elder and 120 Minute IPA. First, participants might not use it at all because they think it is irrelevant. Second, they might shift the 50% consensus distribution to the right until it shows a 90% consensus in favor of Pliny the Elder, leaving its shape largely intact. Such a shift would reduce participants' estimates of the mean, mode, and the probability of the maximum difference in liking compared to participants' estimates in the control condition, but participants would still underestimate the probability of the smallest possible positive difference. Finally, participants might skew the 50% consensus distribution to the right until it shows a 90% consensus in favor of Pliny the Elder, thereby changing its shape but leaving the mode of the distribution close to 0. Such a shift would produce the descending staircase shape depicted in Figure 1, middle row, 94% consensus level, and display learning of the correct distributional shape. If participants did this, all their estimates of the distributional parameters including the probability of the smallest positive difference in liking should become more accurate.

Method

We recruited 300 participants via Prolific to complete our online study in exchange for £0.90 (48% female, M_{age} = 27.63, *SD* = 9.67). All participants were asked to build the distribution of liking differences for the 94% consensus beer pair as in Experiment 3. Participants in the *debias condition* were first told that they would see an example of the task they were about to complete for a different pair—Trois Pistoles and Maudite—rated by users on beeradvocates.com. Participants further learned that the percentage of users preferring each beer is 50%, and were provided with a picture of the beer bottles. They were then asked, "What do you think the distribution of ratings looks like in this case?" Participants read the same instructions for how to use the distribution builder as in Experiment 3, and were told, "Below we have already filled in the correct answer for this pair of beers, once you advance you will complete this task for a different pair of beers with different information." Upon advancing they were then shown the beer-pair with 94% consensus and asked to complete the distribution builder. As in Experiment 3, participants were incentivized with a 20-cent bonus if their distribution matched the true distribution.

Like before, we predicted participants in the *control condition* would overestimate the mean, the mode, and the probability of the maximum difference in liking, and would underestimate the probability of the mode and of the smallest positive difference in liking. We predicted that a majority of participants in the *control condition* would commit each of these errors, but that the proportion doing so in the *debias condition* would be lower. Furthermore, we predicted that those in the *debias condition* would show less of over- and underestimation of the parameters in question.

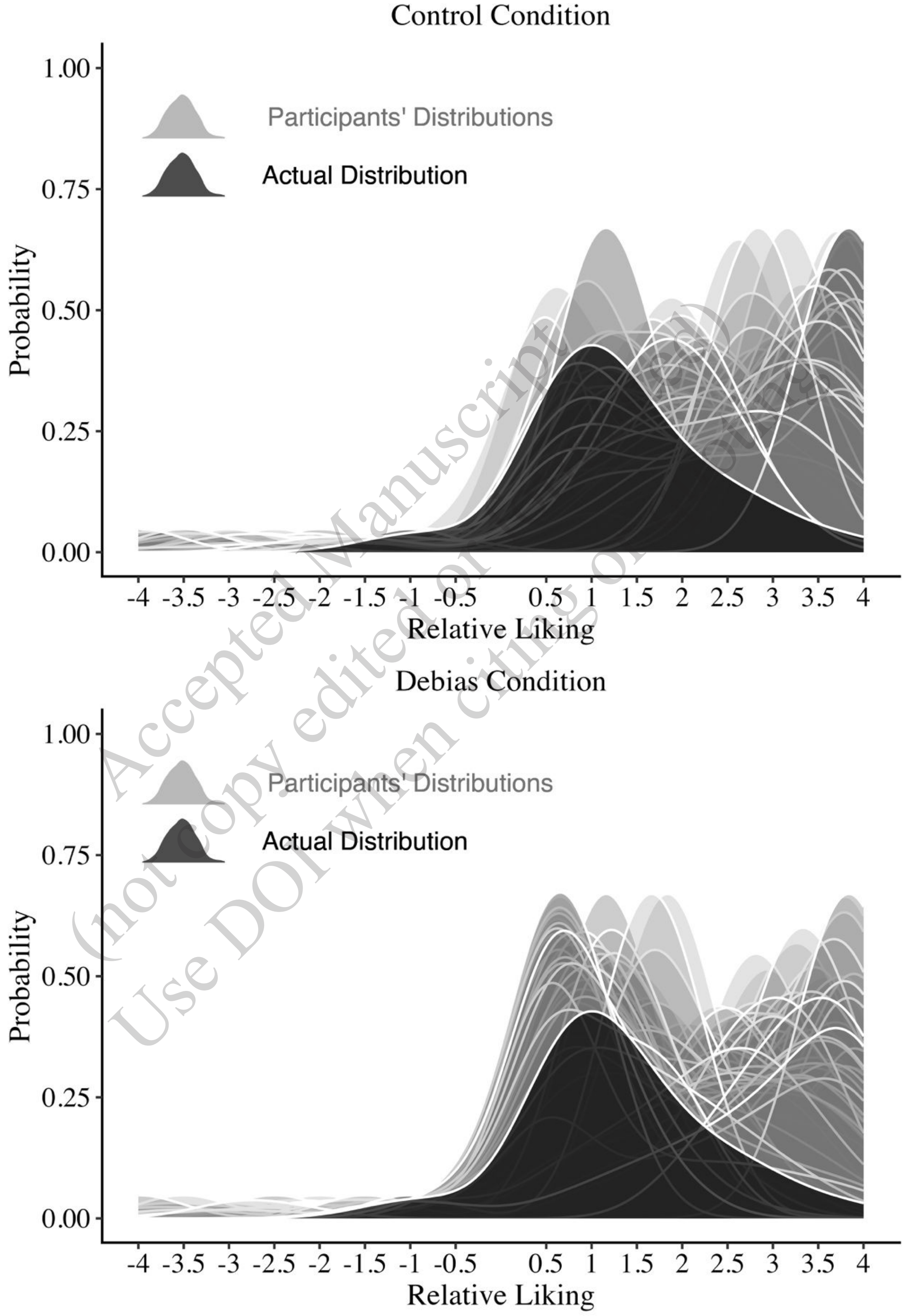
Results and Discussion

A summary of the results is provided in Table 4 (right columns), for the full results see Appendix D Table S2. Figure 7 displays every participant's estimated distribution (in grey) and the actual distribution (in black), for both the *control* and *debias conditions*. We replicate all the findings from Experiment 3 in the *control condition* (all ps < .001). In contrast, the percentage of participants erring in the *debias condition* was reduced by at least 30%, and the degree to which they erred was significantly lessened. All tests comparing the average estimate of the *debias condition* to the *control condition* and the difference in the percentage erring are highly significant (all ps < .001). For the probability estimate of the smallest positive difference, the debiasing treatment removed overestimation entirely.

The effect of debiasing can be seen in Figure 7 as many of the participants' distributions have a significant amount of mass that overlaps with the actual distribution of liking differences, with many more peaks closer to the actual peak (i.e., mode). These results suggest that participants are not merely shifting the referent 50% consensus distribution over to create a faulty shaped distribution, but are adjusting the shape of their distribution, which ultimately more closely corresponds with the shape of the actual distribution of liking differences, a sign of learning the correct distributional shape.

FIGURE 7

ESTIMATED DISTRIBUTION VS ACTUAL DISTRIBUTION IN EXPERIMENT 4



EXPERIMENT 5: THE POWER OF CONSENSUS INFORMATION IN SHAPING CONSUMERS' CHOICES

In Experiments 1-4, we have shown that consumers overestimate how much more the leading option in a poll is liked compared to the runner-up. One implication of this finding is that consensus information may be more persuasive due to this overestimation. If consensus information is more persuasive than seeing information about the average liking of the options, then consumers should be more likely to select a majority preferred option over a minority preferred option when they learn consensus information compared to the average liking of the two options. To test this, in Experiment 5 we asked participants to choose between a more expensive hotel A and a cheaper hotel B. Some participants were told about the proportion of other people preferring hotel A over B (consensus information), while others were told the average ratings for each hotel.

Method

We invited 450 participants via Amazon Mechanical Turk to participate in our study in exchange for $.20. Four hundred and fifty-one MTurkers completed the study (53.4% female, M_{age} = 37.9, *SD* = 12.4). Participants were asked to choose between two hotels A and B—Hotel A costs $125 a night and Hotel B $98 a night—and were further informed that 500 people had rated the two hotels (Hotel A and Hotel B) out of 100 points (1 = low quality, 50 = average quality, 100 = high quality). Participants were then randomly assigned to one of three conditions:

In the *percent consensus* condition, participants learned that “70% of people rated Hotel A higher and 30% of people rated Hotel B higher.” In the *average ratings* condition, participants learned that the average rating of Hotel A is 87 out of 100 and the average rating of Hotel B is 75 out of 100.

Because any given consensus level is consistent with many differences in ratings, we referred to our simulations described in Web Appendix A to derive a corresponding difference. Specifically, using simulations with uniformly distributed liking ratings for two options ranging from 1 to 100, a 70% consensus level is associated with an average difference of 12. Given the skewed nature of the distribution of differences in ratings, this average difference covers more than half of all rating differences (see Figure S1).

Comparing choice shares in the *percent consensus* and the *average ratings* conditions allows us to test the relative persuasiveness of consensus information compared to average ratings. We included a third *out of consensus* condition to test a potential alternative explanation for why a greater proportion of participants in the *percent consensus* condition will choose the more expensive Hotel A. In the third *out of consensus* condition, participants learned that “7 out of 10 people rated Hotel A higher and 3 out of 10 people rated Hotel B higher.” According to the alternative explanation, participants may anchor on the difference between the numbers provided, rather than infer liking differences from consensus information. That difference is 40% in the *percent consensus* condition compared to 12 rating points in the *average ratings* condition, which should produce a larger contrast in favor of the more expensive Hotel A in the former condition (e.g., Fernberger 1920; Heintz 1950; Sherif, Taub, and Hovland 1958; Wever and

Zener 1928). To rule out this alternative account, the third *out of consensus* condition frames consensus information as “7 out of 10 rated Hotel A higher and 3 out of 10 rated Hotel B higher.” The difference in numbers provided here is 4, compared to the 12 rating point difference in the *average ratings* condition. If participants anchor on the difference between the numbers provided, a greater proportion should choose the more expensive Hotel A in the *average ratings* condition than in the *out of consensus* condition. In contrast, we predicted that a greater proportion would choose Hotel A in the *out of consensus* condition than the *average ratings* condition.

Results and Discussion

As predicted, participants were more likely to choose the more expensive Hotel A in the *percent consensus* than the *average ratings* condition (75.7% vs. 42.1%, Pearson $\chi^2(1) = 34.85$, $p < .001$). Contrary to anchoring, participants were also more likely to choose the more expensive Hotel A in the *out of consensus* than the *average ratings* condition (79.5% vs. 42.1%, Pearson $\chi^2(1) = 44.34$, $p < .001$). Choice shares did not differ between *percent consensus* and *out of consensus* conditions (75.7% vs. 79.5%, Pearson $\chi^2(1) = 0.62$, $p = .431$).

Supporting the prediction that consensus information is more persuasive, a larger proportion chose the more expensive Hotel A when informed about the proportion of others preferring it rather than the average ratings of the two hotels. This difference was not caused by participants anchoring on the larger numerical difference of “40%” in the *percent consensus* condition, because we observed the same result when consensus was described as 7 (vs. 3) out of 10 (and hence the numerical difference was only “4”). Importantly, the results of Experiment 5 suggest that communicators can sway consumers’ preferences by choosing how to display information about other consumers’ preferences.

EXPERIMENT 6: INFERRING ESPN EXPERTS' POINT SPREAD PREDICTIONS FOR SUPER BOWL GAMES

Experiment 6 was designed as a test of overestimating differences using real world data in a different domain: prediction polls about a major sports event. Every year, ESPN football commentators and analysts predict the outcome of the Super Bowl. Interestingly, ESPN.com provides readers with the percentage of experts predicting each team to win, that is consensus information. This consensus information is directly calculated from the experts' score predictions for each team, which are provided on the same webpage right below the consensus information. Thus, for each expert we know how much better they think the winning team is than the losing team (i.e., we know each expert's point difference or point spread).

From the four Super Bowl games for which data are available on ESPN.com (2020), we chose Super Bowl LI and LIII. For Super Bowl LIII, 62% of experts predicted Team A to win by, on average, 0.97 points. The Vegas spread was 2.5 points, and Team A actually won the game by 10 points. For Super Bowl LI, 72% of experts predicted Team A to win by, on average, 1.87 points. The Vegas spread was 3 points, and Team A actually won the game by 6 points. In our experiment, we presented participants with the ESPN experts' prediction consensus of either game (in which we replaced the actual names of the teams with Team A and Team B), and asked them to guess the average spread predicted by the experts (estimates were monetarily incentivized for accuracy).

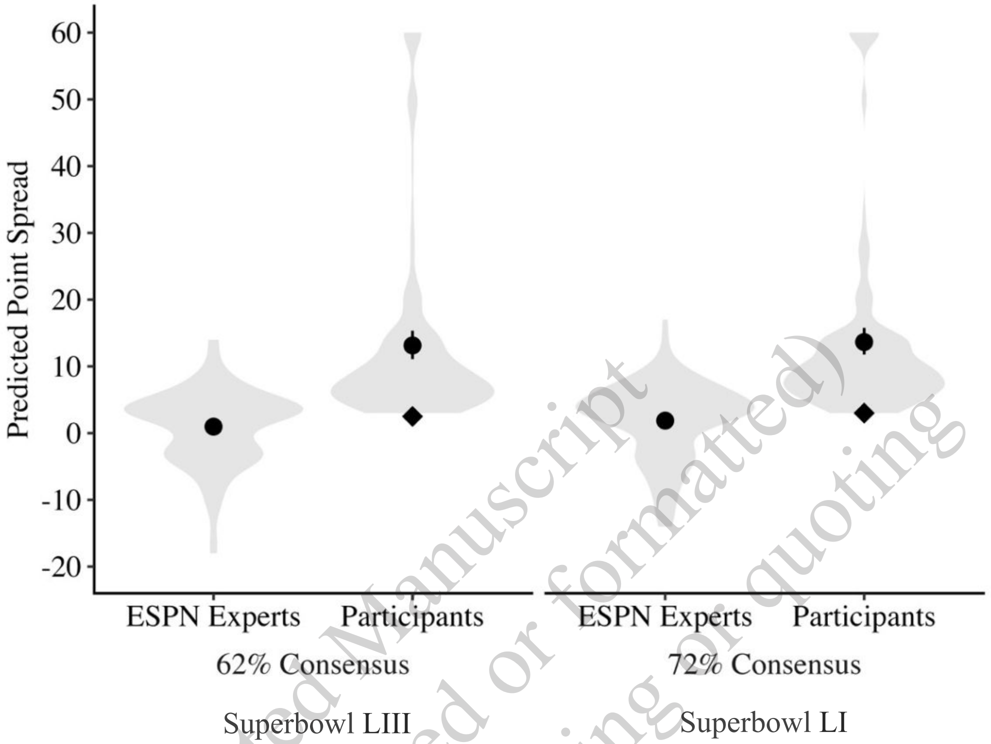
Method

We recruited 400 participants via Amazon Mechanical Turk to complete our study in exchange for $.20 and a potential bonus of $.10 for answers within 10% of the actual point spreads. In our MTurk listing, we requested football fans. To ensure participants were football fans, MTurkers entering our survey were asked two screening questions (number of players on the field, and identification of a referee signal). Participants who failed to answer both questions correctly were (as preregistered) not allowed to continue with our study. Of those who successfully passed this screening, we dropped anyone who took our survey more than once leaving us with 368 participants (79.8% female, M_{age} = 37.2, *SD* = 11).

Participants were randomly assigned to either the Super Bowl LI (72% consensus) or the Super Bowl LIII (62% consensus) condition. In each condition, participants learned that "ESPN experts regularly predict the outcomes of professional football games." We then presented them with information specific to the Super Bowl game to which they had been assigned to. In the 62% [72%] condition, they read "For one such game (we cannot disclose the name of the teams, so we will call them Team A and Team B), 97 [100] ESPN football experts predicted the final scores for each team.[…] 62% [72%] of the ESPN experts predicted Team A would win. 38% [28%] of the ESPN experts predicted Team B would win."

Participants were then asked, "What do you think is the average point spread predicted by the ESPN experts (i.e., the average predicted number of points by which one team wins over the other)?" Participants were informed that if their answer was within 10% of the average point spread of the experts, they would get a $0.10 bonus.

FIGURE 8
RESULTS OF EXPERIMENT 6



NOTE.— Predicted point spreads for Super Bowl games LIII and LI winsorized for participants at the 5th and 95th percentiles. The black diamond marks the Las Vegas point spread. The black circles are the average estimate for that group with error bars showing 95% CIs. The grey cloud shows the distribution.

Results and Discussion

In order to control for outlandish guesses, we winsorized (as preregistered) the data at the 5th and 95th percentiles. As predicted, participants overestimated experts' predicted point spreads, both for the 62% consensus ($M_{\text{predicted}}$ = 13.13, *SD* = 14.14; M_{actual} = 0.97, *SD* = 5.67; $t(273)$ = 8.12, $p < .001$) and the 72% consensus ($M_{\text{predicted}}$ = 13.64, *SD* = 13.85; M_{actual} = 1.87, *SD* = 5.86; $t(288)$ = 8.12, $p < .001$, see Figure 8). These results are robust across stricter winsorization cutoffs and persist when compared to the actual Vegas spreads.

Experiment 6 demonstrates that football fans, who were told the proportion of ESPN experts predicting the winner of two Super Bowl games, overestimate the average point spread predicted by the sports experts. Participants' estimates were on average four times as high as the Vegas spreads, and at least 7.5 times larger than the ESPN experts' point spreads. The results of Experiment 6 thus provide support for our hypothesis in sports predictions, showing that prediction polls about upcoming games can lead consumers to overestimate how much better the leading team is.

EXPERIMENT 7: UNDERESTIMATING AN ELECTION'S WINNER VOTE SHARE

In Experiment 7, we extend the findings of Experiment 6 to a different setting, elections, which has more than two options in the poll. We acquired a dataset from a real poll that asked citizens of a European country which party they would vote for in their upcoming national election, and also asked them to rate the likelihood they would vote for each party. Thus, we had the percentage of polled citizens reporting to vote for each party (consensus information) as well as the average likelihood to vote for each party (liking of each party). These data contain a choice of more than two options, hence, the percentage preferring the majority chosen (the consensus level) option can be less than 50%. Participants in Experiment 7 were either presented with poll consensus information or the average rating of the likelihood to vote for each party from the poll, and asked to predict the actual vote share that the dominant party had received in the election. We predicted that participants presented with the average likelihood to vote ratings would underestimate the vote share of the winning party compared to participants who were presented with the poll consensus information. This is because when presented with average liking ratings, people tend to associate small differences in liking with consensus levels that are too small. Experiment 7 hence tested the flipside of the overestimation prediction that was tested in the previous experiments. We again incentivized participants' accuracy of estimates with a monetary bonus.

Method

We recruited 600 participants via Prolific to complete our study in exchange for $.20 and a potential bonus of $1.00 for the 10 closest answers to the actual vote shares of the winning party (50.5% female, M_{age} = 30.5, *SD* = 9.65).

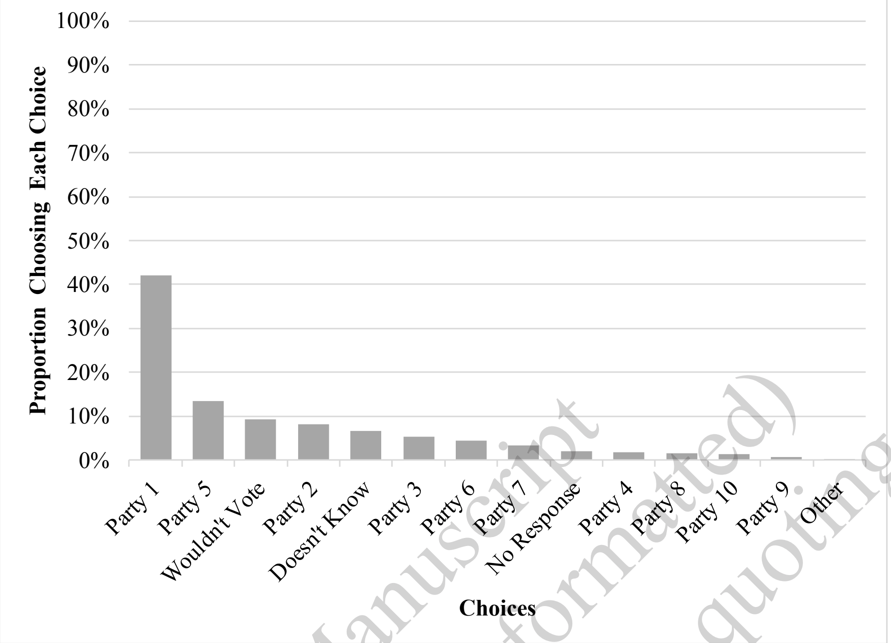
For this study, we used data from a new dataset which contains responses from 1500 citizens in a European country who were polled about an upcoming election for seats in their national government. Due to an NDA, we cannot disclose the country nor the actual dataset beyond mentioning the variables we use and the method of polling. The poll was conducted in-person and citizens were sampled by national quotas of age, gender, city type, and level of education. In addition, datapoints were weighted to correspond to a nationally representative sample. The election consisted of 10 primary parties for which participants could vote. Participants could also write in another party or not select one of the ten parties. Respondents indicated which party they would vote for, and also rated how likely they were to vote for each party on a scale from 0 ("not at all likely") to 10 ("very likely"). From the first vote-choice question we calculated choice-shares for the ten parties (i.e., consensus information). From the ten likelihood ratings we calculated the average likelihood to vote for each party (liking ratings). Not all of the 1500 participants rated their likelihood to vote for all of the 10 parties. Thus, in order to make the information comparable between the likelihood to vote ratings and consensus information, we use only the subset of participants who answered the party preference question and all ten likelihood to vote questions. This subset consists of 1,269 polled citizens.

The six-hundred participants in the Prolific study first learned that we were interested in their thoughts about an election that had already taken place. They then read “The election was in a European country and decided how many seats in the government each party would be allocated. Due to confidentiality, we are not allowed to disclose which country's election it was.” On the next screen they read “Before the election, a poll was conducted asking citizens which of the 10 parties up for election they would vote for. The survey was administered in person and the respondents of the poll form a representative sample from that country.” Participants were then randomly assigned to one of two conditions, consensus information or ratings information.

Participants in the *consensus condition* read, “Before the election, a poll was conducted asking citizens which of the 10 parties up for election they would vote for. The survey was administered in person and the respondents of the poll form a representative sample from that country. The poll asked 1269 citizens to make a single choice indicating which of the 10 parties they would vote for. […]. In the graph below, you see the percentage of polled citizens voting for each party.” The graph, reproduced in Figure 9, displayed that 42% of polled citizens reported that they would vote for Party 1. The next highest percentage was 13.5% of polled citizens who reported they would vote for Party 5.

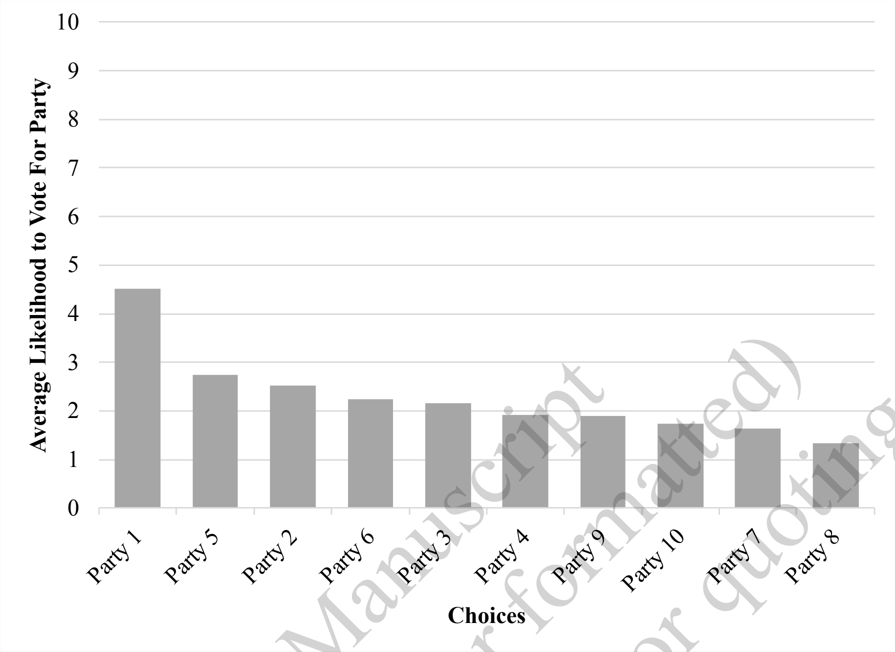
FIGURE 9

GRAPH DISPLAYED IN CONSENSUS CONDITION



In the *ratings condition*, participants read, “Before the election, a poll was conducted asking citizens how likely they would be to vote for each of the 10 parties up for election. The survey was administered in person and the respondents of the poll form a representative sample from that country. The poll asked 1269 citizens how likely they would be to vote for each of the 10 parties on a scale from 0 "Not at all likely" to 10 "Very likely".[…]. In the graph below, you see the average likelihood of voting for each of the 10 parties from the polled citizens.” The graph, reproduced in Figure 10, displayed an average rating of 4.5 out of 10 for Party 1. The next highest average rating was 2.73 out of 10 for Party 5.

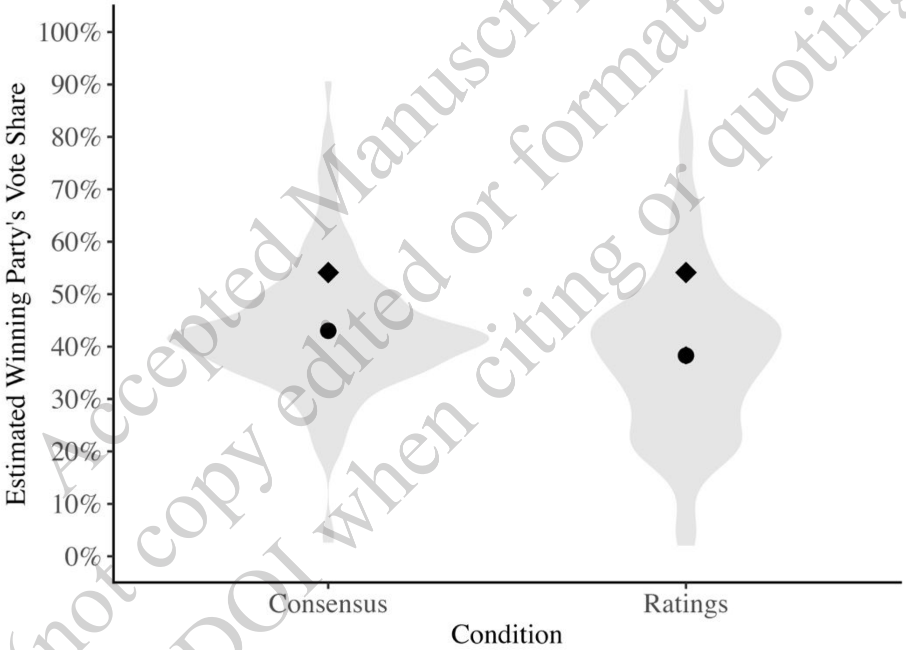
FIGURE 10
GRAPH DISPLAYED IN RATINGS CONDITION



All participants learned, “You are asked to predict the percentage of the total votes that the winning party obtained. We will select 10 winners whose guesses are closest and pay them 1 pound bonus.” After viewing their respective graph, participants saw the main dependent variable and were asked, “What percentage of the total votes do you think Party 1 received in the actual election?” They could enter answers from 0 to 100 with up to two decimal places.

Our key prediction was that participants would estimate the vote share of Party 1 from the actual election (54.13%) to be lower in the *ratings condition* than in the *consensus condition*. Importantly, we note that while the graph for the consensus condition displays a percentage on the y-axis and the graph for the ratings condition displays the average likelihood of polled citizens to vote for a party on the same axis, both graphs have comparable ranges, 0% to 100% and 0 to 10, respectively. Thus, the visual range of their y-axis is identical: 0% to 100% with ticks at every 10% is equivalent to 1 to 10 with ticks at every 1-unit increase. Additionally, recall that the percentage of polled participants selecting they would vote for Party 1 is 42% and the average rating for Party 1 is 4.5. This means that any effect due to the height of the bar for Party 1 is implicitly controlled for. Thus, if participants mapped the objective, numerical information given to them about Party 1—4.5 out of 10 in the *ratings condition* and 42% out of 100% in the *consensus condition*—to the scale of the dependent variable, we should observe a greater predicted vote share in the *ratings condition* than in the *consensus condition*.

FIGURE 11
RESULTS OF EXPERIMENT 7



NOTE.— Estimated winning party’s vote share. The black diamond marks the true vote share received by the winning party, 54.13%. The black circle is the average estimate for that condition. The error bars showing 95% CIs are hard to see because they are not greater than the diameter of the circle.

Results and Discussion

As predicted, participants in the *ratings condition* estimated the actual vote share received by Party 1 to be lower than did participants in the *consensus condition* (M_{ratings} = 38.29, *SD* = 15.42 vs. $M_{\text{consensus}}$ = 43.02, *SD* = 12.40, *t*(598) = 4.13, *p* < .001; see Figure 11).

GENERAL DISCUSSION

In this paper, we examined how consensus information from polls affects consumers' inferences about differences in liking between the poll-choice options. In our empirical investigation, we focused on understanding (1) how consumers draw inferences from consensus information, (2) why these inferences may deviate from reality, (3) how said inferences can impact consumer choices, and (4) how inferences can be debiased. Findings from 6 pre-registered experiments demonstrate that consumers overestimate differences in liking inferred from consensus information (Experiments 1-6). This occurs even when participants have substantial background knowledge on data analytics (Experiment 1), when they have their own subjective beliefs about the choice options (Experiment 2), when they are sophisticated in the judged domain (Experiment 6), and when financial incentives are at stake (Experiments 1, 3, 4, and 5). Our data demonstrate that consensus information causes consumers to overestimate how much better the majority preferred option is than the minority preferred option as they intuitively gravitate toward large differences and neglect small differences (Experiments 3 and 4). These overestimations can lead to shifts in choice shares depending on whether data are displayed in consensus format or as average liking ratings (Experiment 5), and to overestimate by how much a team will win a game (Experiment 6). The flipside, underestimation when inferring vote shares of the leading party in an upcoming election from liking ratings than consensus information, was demonstrated in preregistered Experiment 7. Finally, overestimation is greatly reduced when consumers are shown what the distribution of average differences in liking for a 50% consensus looks like, which causes them to adjust their beliefs about the frequency of large and small differences (Experiment 4).

Theoretical Contributions

The theoretical contribution of our work is threefold. First, while a vast amount of research has examined the psychological processes by which consumers draw inferences about a target (e.g., Broniarczyk and Alba 1994; Dick, Chakravarti, and Biehal 1990; Evangelidis and van Osselaer 2019; Jaccard and Wood 1988; Johnson and Levin 1985; Kardes, Posavac, and Cronley 2004), it has not examined whether said inference processes lead to accurate inferences, estimates, or judgments. In many cases it is difficult—or even impossible—to examine the accuracy of consumers' inferences because there are no normative benchmarks to which said inferences can be compared. Our settings allow for such comparisons because it is possible to compare estimates of differences in liking to observed distributions of real-world liking ratings, or to simulate a normative benchmark—the likelihood distribution of liking differences—for a given consensus level. Our work thus extends prior research on decision-makers' inferences by demonstrating a robust bias that arises from inference-making processes.

Second, we contribute to the understanding of social influence, specifically of social proof (Cialdini 2007). Social proof denotes the phenomenon that consumers conform to the behavior of others. For instance, hotel guests are more likely to reuse their towels when learning that a majority of other hotel guests do so (Goldstein, Cialdini and Griskevicius 2008). As the name "social proof" indicates, such conforming behavior is explained by consumers' beliefs that others may have more accurate preferences or superior information (Burnkrant and Cousineau 1975; Kelley 1967), especially in ambiguous situations in which they are uncertain about their own preferences (Young et al. 2014). Since consumers conform out of a belief that others' responses provide diagnostic information, social proof often leads to private acceptance as well as public compliance (Cialdini 2007). Consensus information in polls apprises consumers of the option being preferred by the majority, so it constitutes a form of social proof (Cialdini 2007). Our findings hence suggest an additional explanation why social proof is so powerful in making consumers conform to the preferences and choices of others. Consumers not only believe others' preferences to be diagnostic, they are also likely to overestimate others' strength of preferences (i.e., differences in liking of the options), thereby wrongly inferring that the majority preferred option is much better than it actually is, or inferring that the minority preferred option is much worse than it actually is. Thus, social proof consists not only of conforming to the preferences and choices of others, it also involves exaggeration of others' liking differences.

Our third theoretical contribution concerns the consequences of our findings for preference inferences over time. Conforming to the preferences of others over time can result in self-reinforcing social influence effects whereby the number of consumers whose preferences are mimicked grows with the number of consumers conforming to others' preferences, leading to so-called "preference cascades." Investors, for example, are more likely to follow other investors' coverage of a firm as the number of investors following that firm increases (Rao, Greve, and Davis 2001), and higher levels of scarcity of a good can increase further sales of the scarce good (Banerjee 1992; Van Herpen, Pieters, and Zeelenberg 2009). According to our findings, exaggeration of social influence effects not only happen over time as in the case of preference cascades, they can also occur instantaneously as consumers exaggerate the preferences to which they conform. Preference cascades, in turn, may be further accelerated by consumers' exaggeration of the preferences that they conform to.

Managerial and Public Policy Implications

As the results of Experiments 5, 6 and 7 show, managers, politicians, public policy makers—in short anyone interested in influencing consumers' judgments and choices—can strategically choose how to display aggregate preferences. In many cases, displaying others' preferences as consensus rather than average differences in liking will make the majority-preferred option more attractive, and hence increase its choice share. A company engaging in time-consuming data collection by eliciting ratings of their own and competitor products may instead conduct simple polls. Poll results (consensus information) will in many cases be more persuasive than more elaborate average ratings. As mentioned previously, Pepsi has been using this strategy successfully for the last 30 years to gain market share from its main rival.

Consumers, on the other hand, should be wary of the inferences they draw from consensus information and seek out more continuous measures that can help to better inform their choices. This is applicable to any domain in which consumers infer differences on some variable from consensus information, such as polls about public opinion concerning political candidates (e.g., "Which candidate do you trust more to revive the economy?"), sports contenders ("Who will win the Super Bowl?"), the next best product to create or keep ("Cast your DEWcision for the flavor to keep on shelves."; prnewswire.com 2016), or product choices ("78% prefer our product over our competitor's product"). As consumers tend to form unrealistic expectations about political candidates, athletes/teams, and products when learning consensus information, they are likely to be disappointed once their favorite candidate is voted into office, they have cast a bet on their preferred contender, or have bought their preferred product.

There are, however, situations in which consensus is actually more informative than knowledge of the average liking of the options. For example, in the 2020 US democratic primary, democrats were trying to figure out which candidate, Bernie Sanders or Joe Biden, would have better chances of beating Donald Trump in the presidential election. Sanders had a minority of Democrats passionately supporting him and feeling very lukewarm about Biden. Biden had the majority of Democrats' support, but most were not very passionate about his candidacy. Chances of beating Trump in the presidential election are here indicated by consensus information, not by the strength of liking for each candidate. In general, which format for expressing aggregate consumer preferences is more advantageous depends on the goal at hand. Probability of superiority is indicated by consensus information, whereas satisfaction with and betting on actual outcomes is better predicted by continuous liking measures.

Implications for Preference Elicitation

Researchers use a variety of methods to elicit preferences, such as choice, preference ratings, or WTP. In doing so, they need to be aware of the kind of information that is acquired in the aggregate. When measuring consumer preferences for two options through choice, researchers learn about the ordinal preference ranking of individuals, and through aggregation obtain ratio-scaled consensus information. When assessing liking through ratings, researchers learn about individual differences in liking, and through aggregation obtain average interval-scaled differences in liking. Note that the two convey different information. Consensus cannot be inferred from aggregate average liking ratings (even when the skew of both distributions is known). To infer consensus, a researcher needs to have access to the individual ratings by each consumer. Likewise, average ratings cannot be inferred from consensus information, only distributions of average differences in liking can, as we have shown in this paper.

In choice experiments, it is quite common for researchers to test choice shares against 50%. In case of a null result, many researchers state that respondents are indifferent between the choice options. As we have demonstrated this is an unwarranted inference because indifference—a liking difference—cannot be inferred from consensus information. A 50% choice share may indicate indifference, but may also be indicative of uniform or bipolar distributions of liking. In fact, any symmetrical liking distribution will lead to a 50%-50% choice share[4]. Consider the US presidential election in 2016. It would have been foolish to infer from a poll showing 52% of voters preferring Trump over Clinton that the average voter likes Trump as much as Clinton.

Conclusion

Polls are everywhere. They inform consumers of others' beliefs and preferences across a variety of consequential domains from beliefs in climate change to big purchase decisions like cars. Consumers relying on such polls to inform themselves of what to think or do are faced with the problem of having to infer strength of beliefs and liking from consensus. We find that consumers often overestimate how much better the leading option in a poll is, and in doing so risk making decisions that they may regret.

[4] We thank Irene Scopelliti for pointing out this application of our findings.

DATA COLLECTION STATEMENT

The first author collected data for all experiments except Experiment 1, which was collected by Professor Ioannis Evangelidis. Experiment 1 was administered via Qualtrics to students in a data analytics course at ESADE in March of 2021. The first author analyzed all data reported in this paper. Data were collected between 2019 and 2023 on Amazon's Mechanical Turk, ESADE class, and on Prolific. All data, materials and preregistrations can be accessed at https://researchbox.org/446.

REFERENCES

Autoguide.com (2017). url: https://www.autoguide.com/auto-news/2017/09/poll-toyota-camry-or-honda-accord-.html. accessed on 6/18/2020.

Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, *107*, 797-817.

Breitbart.com (2021). url: https://www.breitbart.com/politics/2021/02/16/poll-donald-trump-favorability-75-among-republicans-mitch-mcconnell-underwater-at-15/. accessed on 10/20/2021.

Broniarczyk, S. M., & Alba, J. W. (1994). Theory versus data in prediction and correlation tasks. *Organizational Behavior and Human Decision Processes*, (1), 117-139.

Burnkrant, R. E., & Cousineau, A. (1975). Informational and normative social influence in buyer behavior. *Journal of Consumer Research*, *2*, 206-215.

Cialdini, R. B., & Cialdini, R. B. (2007). *Influence: The Psychology of Persuasion* (Vol. 55, p. 339). New York: Collins.

Dick, A., Chakravarti, D., & Biehal, G. (1990). Memory-based inferences during consumer choice. *Journal of Consumer Research*, *17*, 82-93.

ESPN (2020). url: https://www.espn.com/nfl/story/_/id/28576384/super-bowl-liv-score-predictions-espn-experts-pick-49ers-chiefs. accessed on 6/18/2020.

Evangelidis, I., & Van Osselaer, S. M. (2018). Points of (dis) parity: Expectation disconfirmation from common attributes in consumer choice. *Journal of Marketing Research*, 55, 1-13.

——— and ——— (2019). Interattribute Evaluation Theory. *Journal of Experimental Psychology: General*, 148, 1733-1746.

Fernberger, S. W. (1920). Interdependence of judgments within the series for the method of constant stimuli. *Journal of Experimental Psychology*, *3*, 126.

Fisher, M., & Keil, F. C. (2018). The binary bias: A systematic distortion in the integration of information. *Psychological Science*, *29*, 1846-1858.

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing stars: How the binary bias distorts the interpretation of customer ratings. *Journal of Consumer Research*, *45*, 471-489.

Forbes (2021). url: https://www.forbes.com/sites/alisondurkee/2021/06/25/majority-of-americans-support-abortion-poll-finds---but-not-later-in-the-pregnancy/?sh=24f9d2fb5074 . accessed on 11/3/2021.

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, *4*, 133-151.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*, 472-482.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*, 1.

Harper, Maxwell F. and Joseph A. Konstan (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 4: 19:1–19:19. <https://doi.org/10.1145/2827872>

Heintz, R. K. (1950). The effect of remote anchoring points upon the judgment of lifted weights. *Journal of Experimental Psychology*, *40*, 584.

History.com (2020). url: https://www.history.com/news/cola-wars-pepsi-new-coke-failure . accessed on 10/07/2020.

Huber, J., Ariely, D., & Fischer, G. (2002). Expressing preferences in a principal-agent task: A comparison of choice, rating, and matching. *Organizational Behavior and Human Decision Processes*, *87*, 66-90.

Jaccard, J., & Wood, G. (1988). The effects of incomplete information on the formation of attitudes toward behavioral alternatives. *Journal of Personality and Social Psychology*, *54*, 580.

Johnson, R. D., & Levin, I. P. (1985). More than meets the eye: The effect of missing information on purchase evaluations. *Journal of Consumer Research*, *12*, 169-177.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*, 697.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Psychology for behavioral economics. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 49-81.

Kardes, F. R., Posavac, S. S., & Cronley, M. L. (2004). Consumer inference: A review of processes, bases, and judgment contexts. *Journal of Consumer Psychology*, *14*, 230-256.

Keeney, R. L., Raiffa, H., & Meyer, R. F. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge University Press.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska Symposium on Motivation*. University of Nebraska Press.

Kupor, D., & Laurin, K. (2020). Probable cause: The influence of prior probabilities on forecasts and perceptions of magnitude. *Journal of Consumer Research*, *46* (5), 833-852.Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215-233.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14* (10), 435-440.

Myers, T. A., Maibach, E., Peters, E., & Leiserowitz, A. (2015). Simple messages help set the record straight about scientific agreement on human-caused climate change: The results of two experiments. *PloS one*, *10* (3), e0120985.

Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*S, 85.

Pepsimax.ie (2019). url: https://pepsimax.ie/. accessed on 10/07/2020.

Phillymag.com (2018). url: https://www.phillymag.com/foobooz/2018/01/23/vote-best-cheesesteak-2018/. accessed on 6/18/2020.

Prnewswire.com (2016). url: https://www.prnewswire.com/news-releases/dear-dew-nation-dewcision-2016-gives-you-the-power-to-choose-between-two-legendary-dew-flavors-300255397.html. accessed on 10/07/2020.

Quentinandre.net (2022). url: https://quentinandre.net. DOI: 10.5281/zenodo.166736, accessed on: 10/26/2022.

Rao, H., Greve, H. R., & Davis, G. F. (2001). Fool's gold: Social proof in the initiation and abandonment of coverage by Wall Street analysts. *Administrative Science Quarterly*, *46*, 502-526.

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research*, *57* (5), 853-877.

Sharpe, W. F., Goldstein, D. G., & Blythe, P. W. (2000). The distribution builder: A tool for inferring investor preferences. *preprint*.

Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology*, *55*, 150.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185* (4157), 1124-1131.

Van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS One*, *10* (2), e0118489.

Van Herpen, E., Pieters, R., & Zeelenberg, M. (2009). When demand accelerates demand: Trailing the bandwagon. *Journal of Consumer Psychology*, *19*, 302-312.

Westwood, S. J., Messing, S., & Lelkes, Y. (2020). Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *The Journal of Politics*, *82* (4), 1530-1544.

Wever, E. G., & Zener, K. E. (1928). The method of absolute judgment in psychophysics. *Psychological Review*, *35*, 466.

HEADINGS LIST

1) THEORETICAL DEVELOPMENT

2) Polls, Choices, and Differences in Liking

2) Consensus Levels and Corresponding Differences in Liking

2) Are Small Differences in Liking Always more Likely than Large Differences?

2) Empirical Demonstration: Small Differences in Liking are More Likely than Large Differences

2) Given Consensus Information, Consumers Overestimate Average Differences in Liking

2) Effects of Larger Versus Smaller Magnitudes

2) Empirical Overview (Experiments 1-7)

1) EXPERIMENT 1: OVERESTIMATION OF AVERAGE DIFFERENCES WITH SIMULATED DATA

2) Method

2) Results and Discussion

1) EXPERIMENT 2: INFERRING RATINGS AND THE ROLE OF PRIORS

2) Method

3) *Stage One*

3) Stage One Results

3) Stage Two

2) Stage Two Results and Discussion

1) EXPERIMENT 3: ELICITING DISTRIBUTIONAL BELIEFS ABOUT DIFFERENCES IN LIKING

2) Method

2) Results and Discussion

2) Counterfactual Sensitivity Analysis regarding Stimulus Sampling

1) EXPERIMENT 4: DEBIASING RESPONDENTS

2) Method

2) Results and Discussion

1) EXPERIMENT 5: THE POWER OF CONSENSUS INFORMATION IN SHAPING CONSUMERS' CHOICES

2) Method

2) Results and Discussion

1) EXPERIMENT 6: INFERRING ESPN EXPERTS' POINT SPREAD PREDICTIONS FOR SUPER BOWL GAMES

2) Method

2) Results and Discussion

1) EXPERIMENT 7: UNDERESTIMATING AN ELECTION'S WINNER VOTE SHARE

2) Method

2) Results and Discussion

2) Counterfactual Sensitivity Analysis regarding Stimulus Sampling

1) GENERAL DISCUSSION

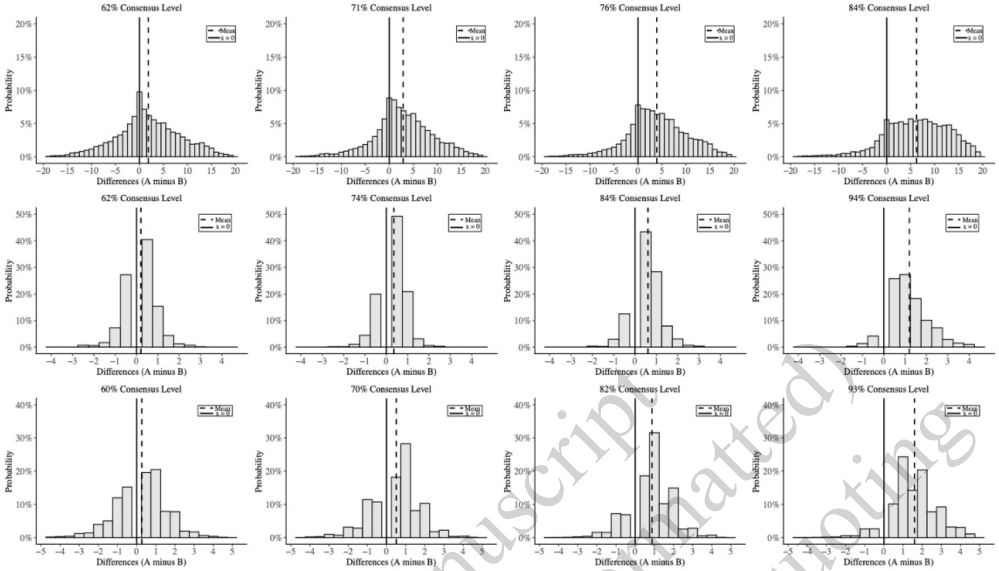
2) Theoretical Contributions

2) Managerial and Public Policy Implications

2) Implications for Preference Elicitation

2) Conclusion

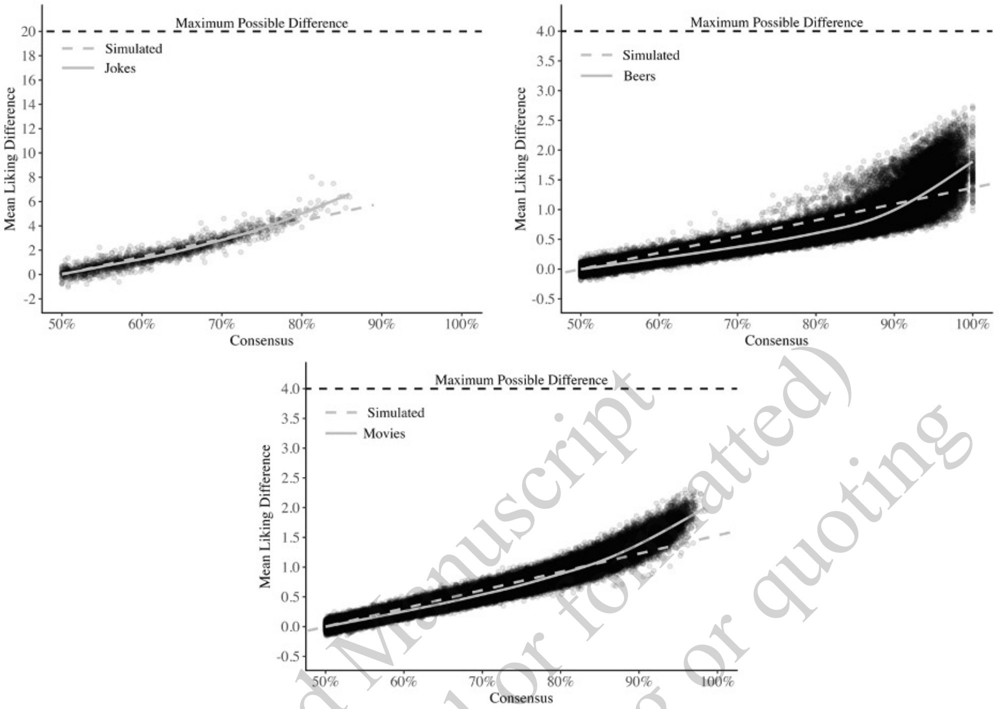
1) REFERENCES



Title: HISTOGRAMS OF RATING DIFFERENCES FOR JOKES (TOP), BEERS (MIDDLE), AND MOVIES (BOTTOM)

NOTE.— Figure 1 shows the distribution of rating differences for four different joke, beer, and movie pairs each, representing low to high consensus levels.

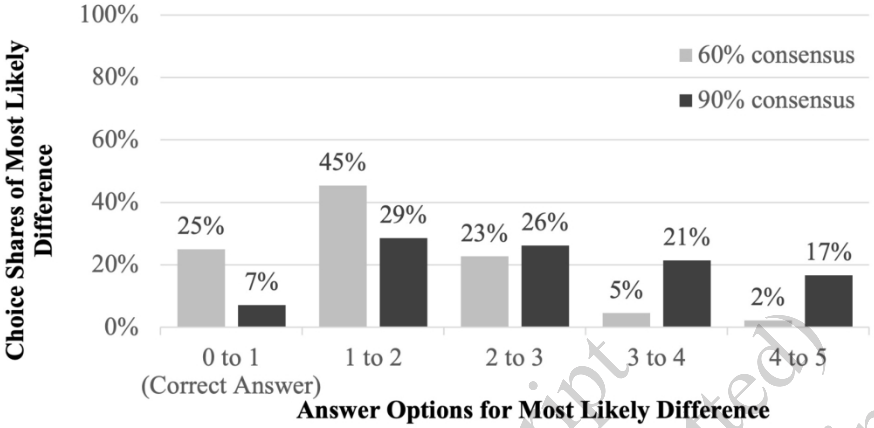
254x145mm (144 x 144 DPI)



Title: SIMULATED AND ACTUAL MEAN DIFFERENCES IN RATINGS

NOTE.— Scatterplots of mean differences in liking for pairs of jokes, beers, and movies, on their respective scales. Darker areas indicate greater density of pairs with the same mean difference. The grey dashed lines display simulated mean differences, and the grey solid lines are the (actual) observed mean differences. The black dashed line marks the maximum possible difference that could occur on the scale for that dataset.

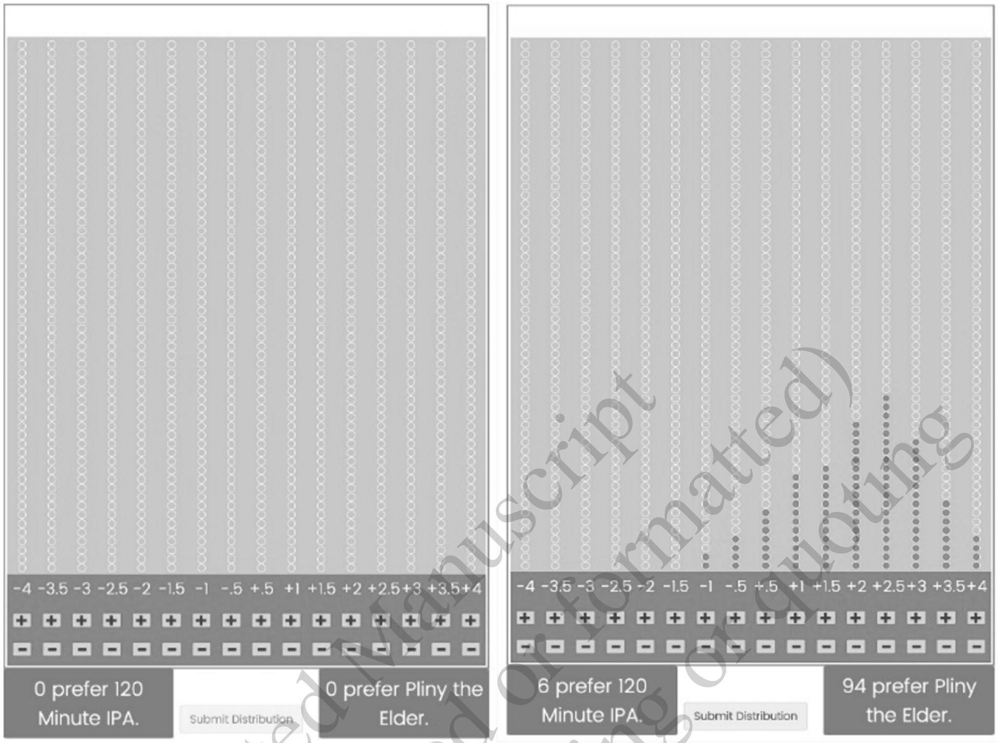
164x116mm (144 x 144 DPI)



Title: RESULTS OF EXPERIMENT 1

NOTE.— Proportion of respondents choosing each answer option for the most likely difference in ratings in Experiment 1.

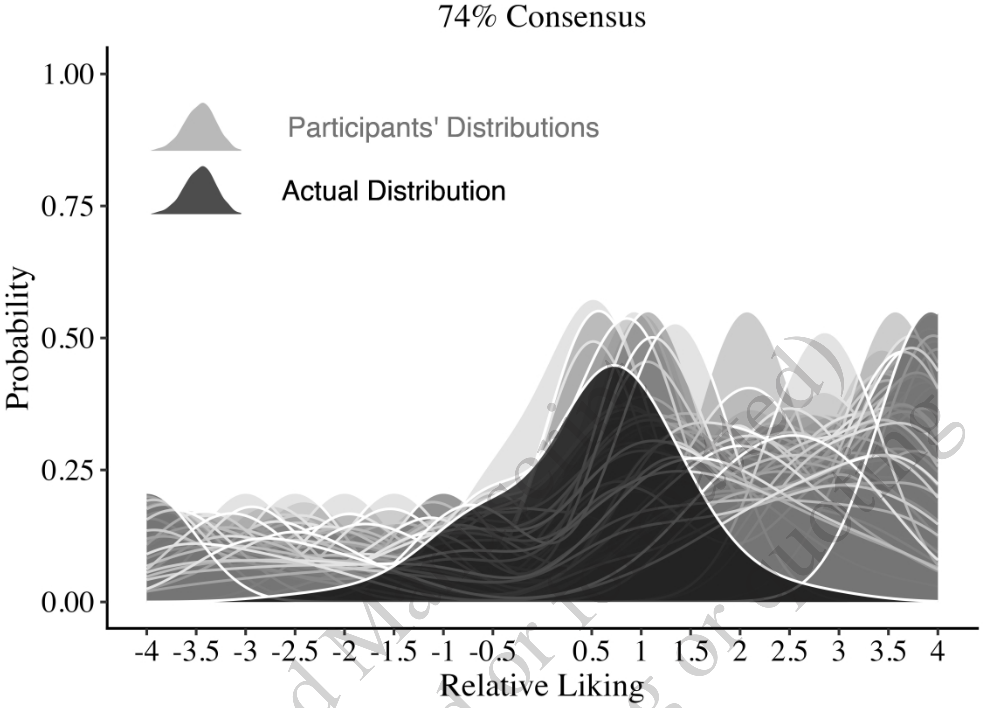
165x76mm (300 x 300 DPI)



Title: EXAMPLE OF DISTRIBUTION BUILDER TASK IN EXPERIMENT 3

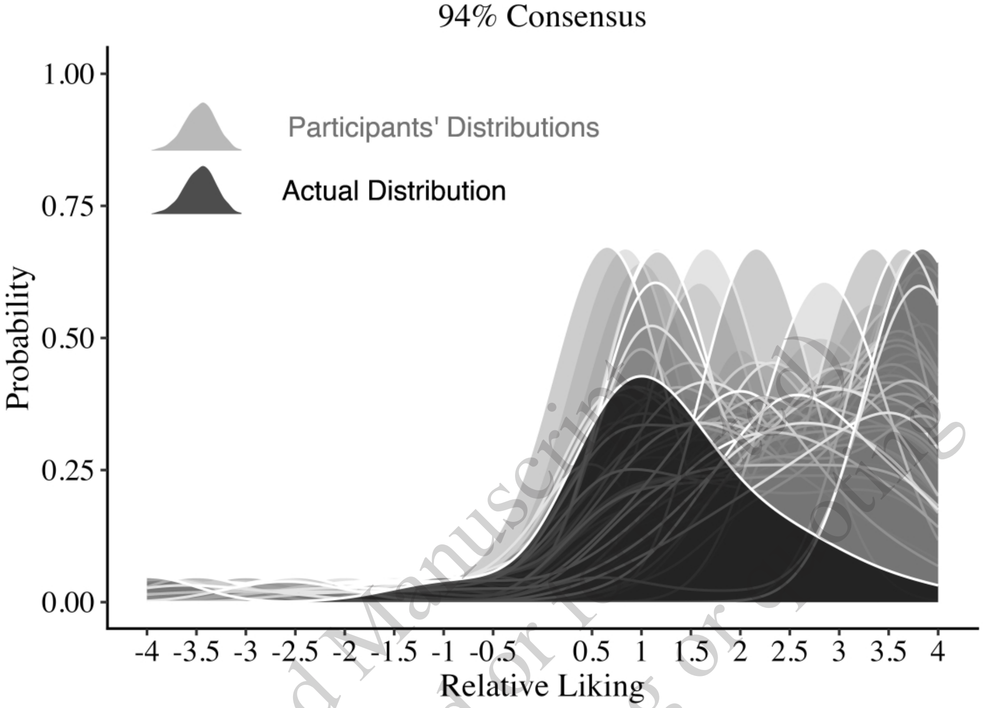
NOTE.— Examples of the distribution builder tool used in Experiments 3 and 4. The left panel is participants' starting point and the right panel represents a participant's final allocation.

204x151mm (144 x 144 DPI)

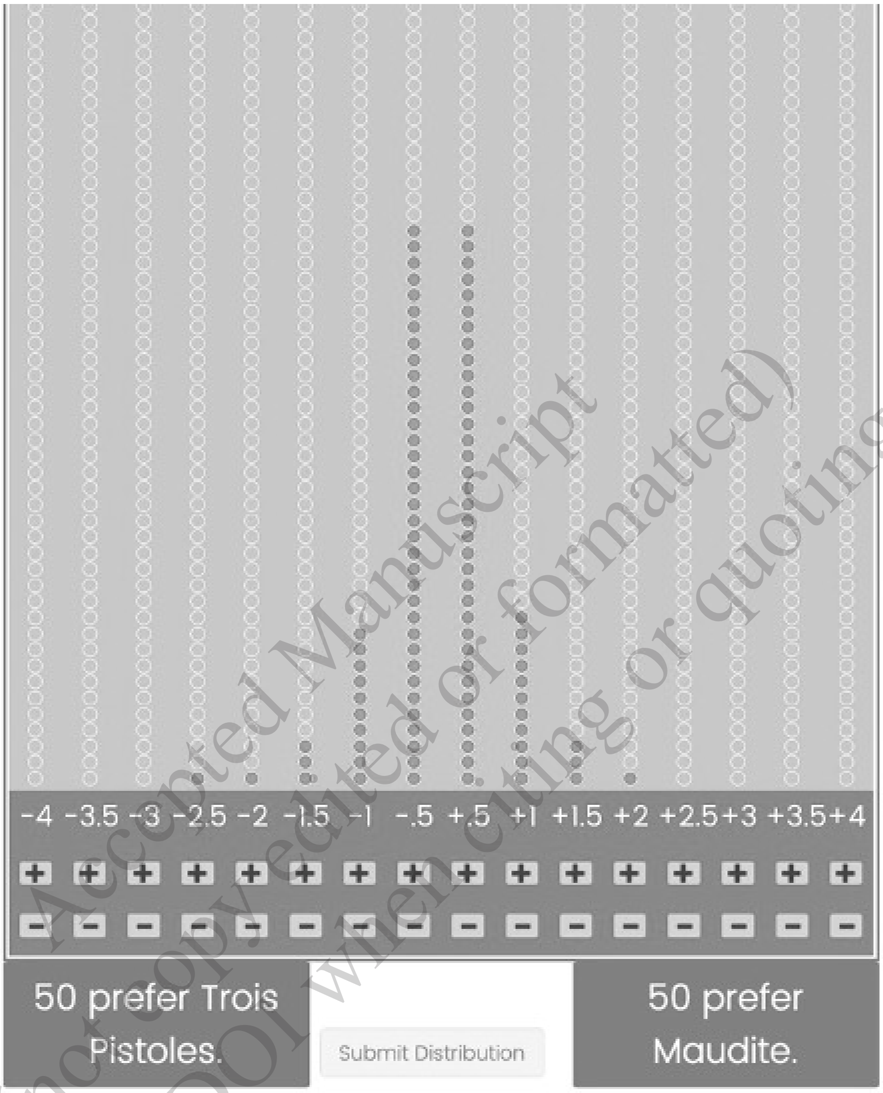


Title: ESTIMATED DISTRIBUTION VS ACTUAL DISTRIBUTION IN EXPERIMENT 3

1940x1411mm (72 x 72 DPI)

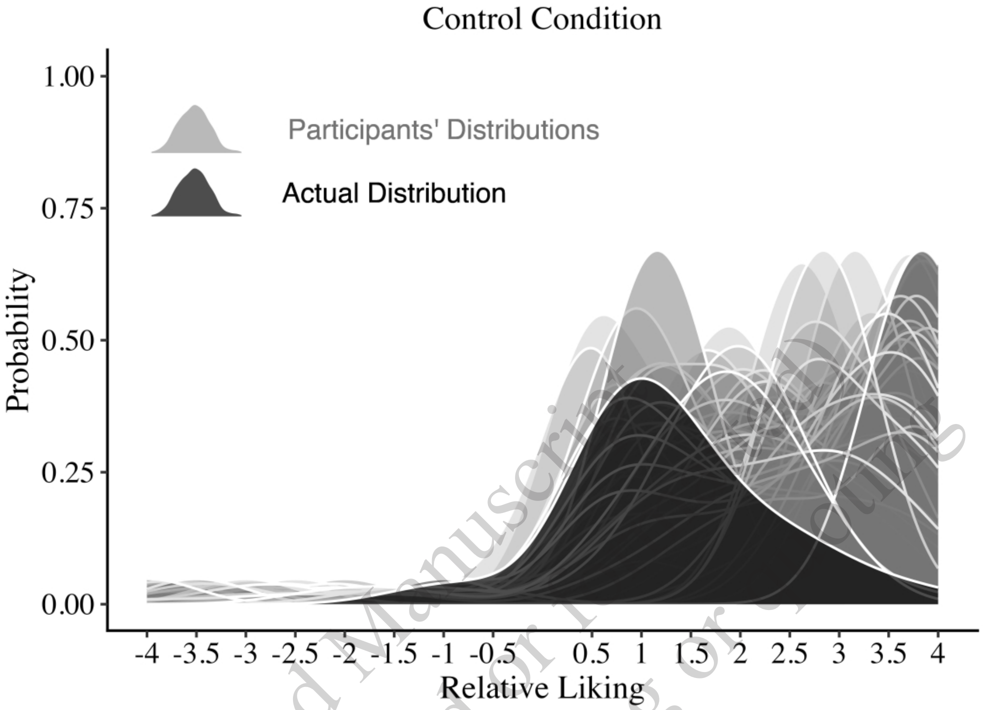


1940x1411mm (72 x 72 DPI)



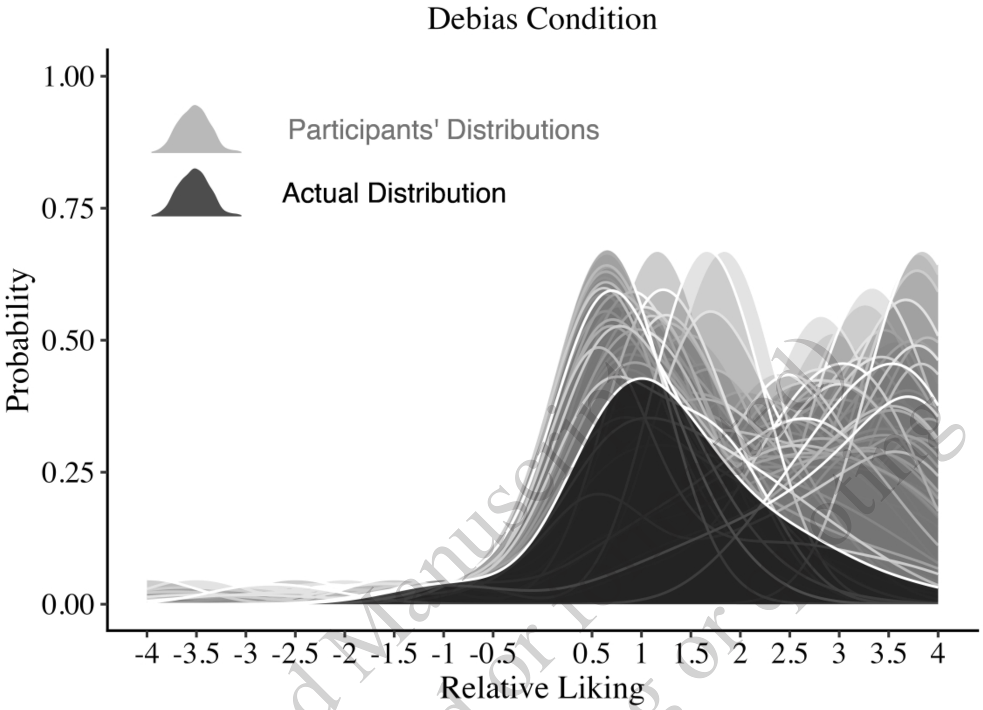
Title: THE 50% CONSENSUS DISTRIBUTION DISPLAYED FOR HALF OF THE PARTICIPANTS IN EXPERIMENT 4

79x96mm (144 x 144 DPI)

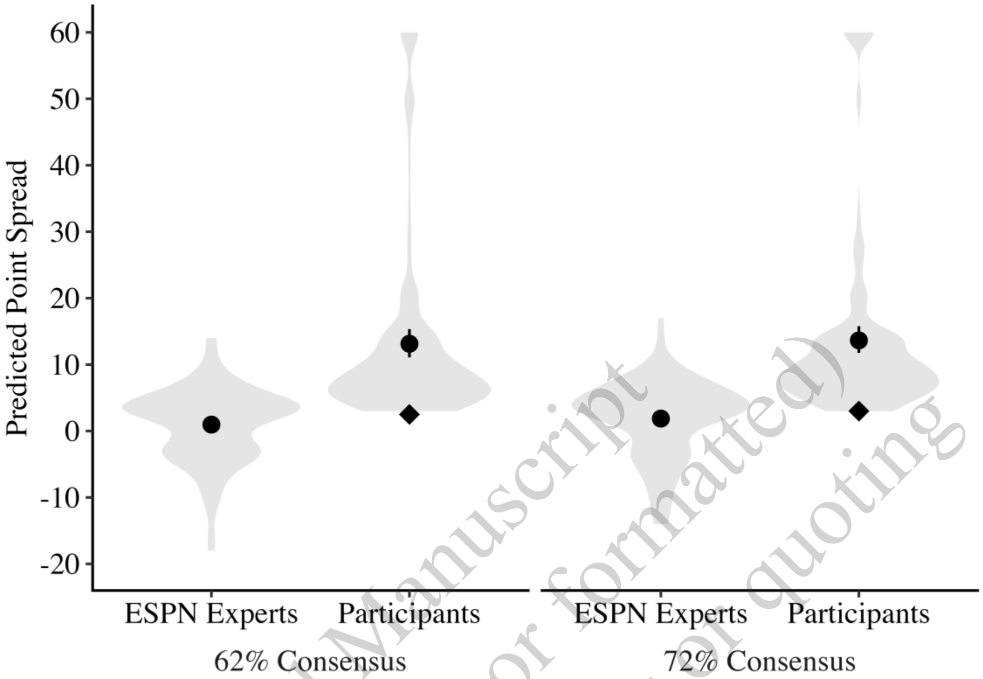


Title: ESTIMATED DISTRIBUTION VS ACTUAL DISTRIBUTION IN EXPERIMENT 4

1940x1411mm (72 x 72 DPI)



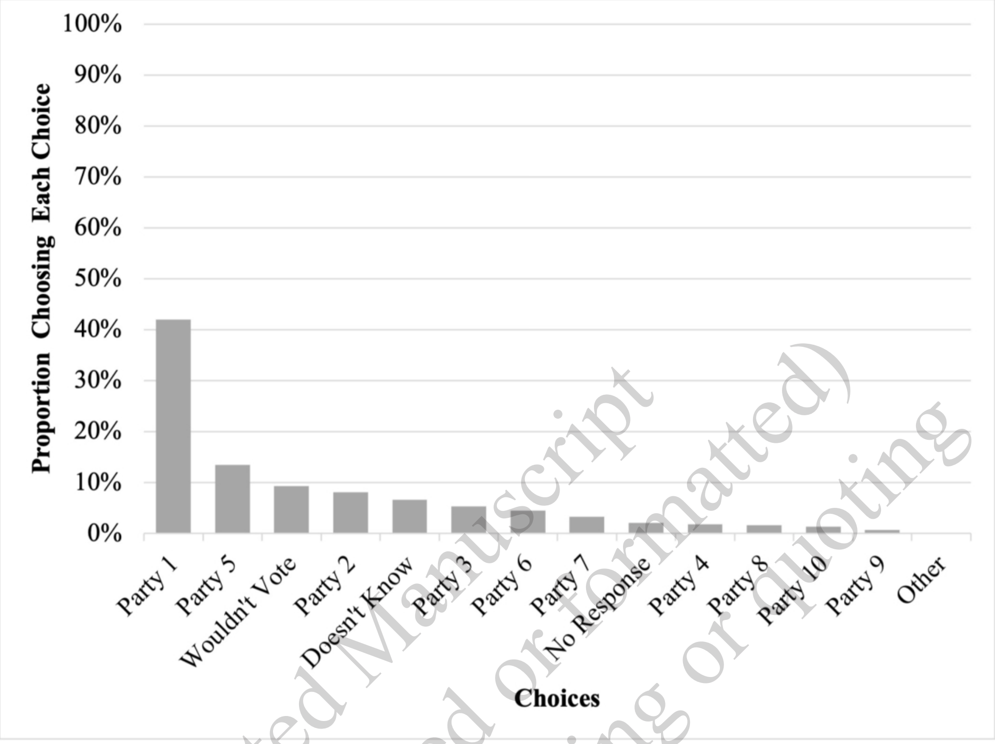
1940x1411mm (72 x 72 DPI)



Title: RESULTS OF EXPERIMENT 6

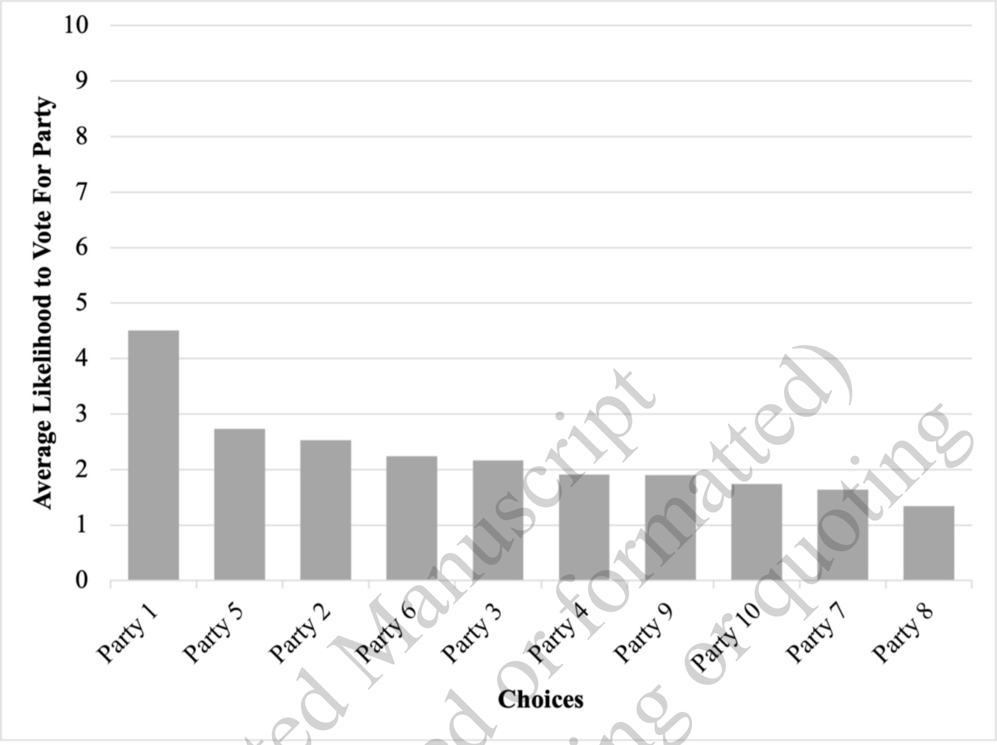
NOTE.— Predicted point spreads for Super Bowl games LIII and LI winsorized for participants at the 5th and 95th percentiles. The black diamond marks the Las Vegas point spread. The black circles are the average estimate for that group with error bars showing 95% CIs. The grey cloud shows the distribution.

1940x1411mm (72 x 72 DPI)



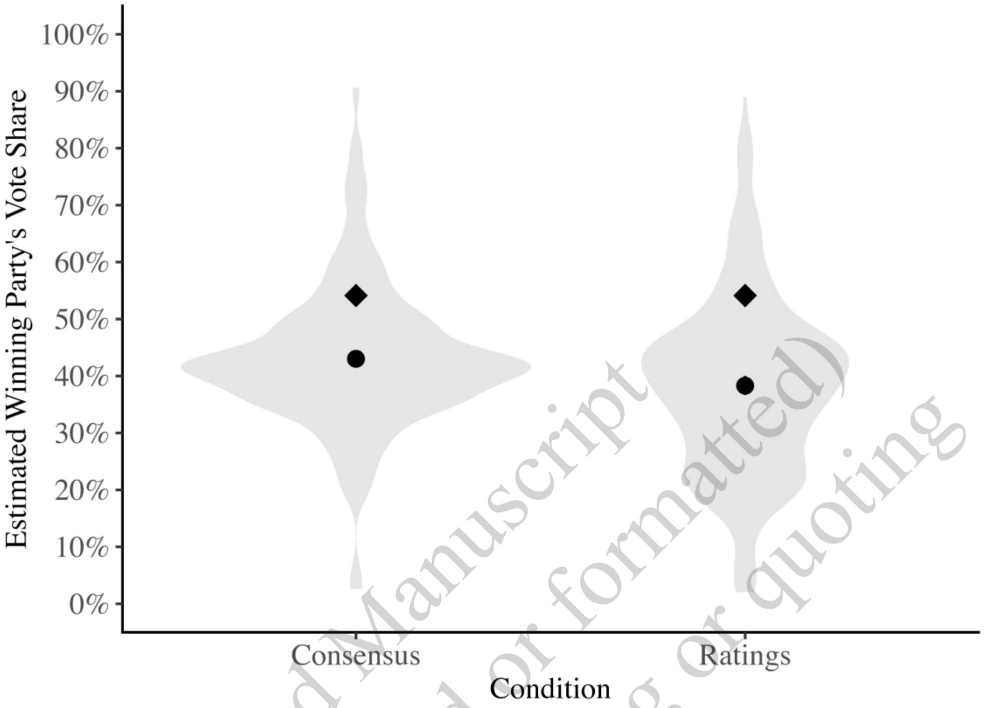
Title: GRAPH DISPLAYED IN CONSENSUS CONDITION

163x121mm (300 x 300 DPI)



Title: GRAPH DISPLAYED IN RATINGS CONDITION

163x121mm (300 x 300 DPI)



Title: RESULTS OF EXPERIMENT 7

NOTE.— Estimated winning party's vote share. The black diamond marks the true vote share received by the winning party, 54.13%. The black circle is the average estimate for that condition. The error bars showing 95% CIs are hard to see because they are not greater than the diameter of the circle.

1940x1411mm (72 x 72 DPI)