Supplementary Materials for:

Does Constructing A Belief Distribution Truly Reduce Overconfidence?

Beidi Hu

University of Pennsylvania

Joseph P. Simmons

University of Pennsylvania

Table of Contents

Supplement 1: Exclusions and Reported Sample Sizes

Table S1 provides the full breakdown of the target sample sizes, pre-registered exclusions, and reported sample sizes for Studies 1-10, S1-S5.

Table S1. Exclusions in Studies 1-S5

Study 1. Target sample size: N = 600
Starting N = 644

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Confidence Interval condition	# excluded from Belief Distribution condition
dropped out of the survey before being assigned to a condition	40	604		
dropped out of the survey after being assigned to a condition	0	604	0	0
were associated with a duplicate IP Address	19	585	10	9
were associated with a duplicate worker ID	2	583	1	1
misreported the worker ID	6	577	4	2
failed the attention check	66	511	27	39

Final sample: N = 511 (262 in the Confidence Interval condition; 249 in the Belief Distribution condition)

Study 2. Target sample size: N = 1,000
Starting N = 1,222

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Confidence Interval condition	# excluded from Belief Distribution condition	# excluded from Control condition
dropped out of the survey before being assigned to a condition	230	992			
dropped out of the survey after being assigned to a condition	0	992	0	0	0
were associated with a duplicate IP Address	33	959	11	8	14
were associated with a duplicate worker ID	11	948	5	2	4
misreported the worker ID	4	944	0	2	2
failed the attention check	132	812	36	53	43

Final sample: N = 812 (278 in the Confidence Interval condition; 263 in the Belief Distribution condition; 271 in the Control condition)
Final sample who met the screening criteria: N = 583 (203 in the Confidence Interval condition; 187 in the Belief Distribution condition; 193 in the Control condition)

Study 3. Target sample size: N = 1,700
Starting N = 2,145

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Confidence Interval / Best Estimate First condition	# excluded from Belief Distribution / Best Estimate First condition	# excluded from Confidence Interval / Best Estimate Last condition	# excluded from Belief Distribution / Best Estimate Last condition	# excluded from Control condition
dropped out of the survey before the attention check	153	1,992					
failed the attention check	17	1,975					
dropped out of the survey after being assigned to a condition	0	1,975	0	0	0	0	0
were associated with a duplicate IP Address	132	1,843	25	23	27	24	33
were associated with a duplicate worker ID	15	1,828	2	3	4	1	5
misreported the worker ID	12	1,816	3	5	1	2	1

Final sample: N = 1,816 (367 in the Confidence Interval / Best Estimate First condition; 365 in the Belief Distribution / Best Estimate First condition; 360 in the Confidence Interval / Best Estimate Last condition; 363 in the Belief Distribution / Best Estimate Last condition; 361 in the Control condition)

Study 4. Target sample size: N = 1,300
Starting N = 1,463

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Belief Distribution / High Precision condition	# excluded from Control / High Precision condition	# excluded from Belief Distribution / Low Precision condition	# excluded from Control / Low Precision condition
dropped out of the survey before the attention check	40	1,423				
failed the attention check	123	1,300				
dropped out of the survey after being assigned to a condition	0	1,300	0	0	0	0
were associated with a duplicate IP Address	80	1,220	16	27	18	19
were associated with a duplicate worker ID	1	1,219	0	0	1	0
misreported the worker ID	6	1,213	2	2	1	1

Final sample: N = 1,213 (299 in the Belief Distribution / High Precision condition; 305 in the Control / High Precision condition; 294 in the Belief Distribution / Low Precision condition; 315 in the Control / Low Precision condition)

Study 5. Target sample size: N = 1,300
Starting N = 1,332

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Belief Distribution condition	# excluded from Control condition
dropped out of the survey before the attention check	10	1,322		
failed the attention check	17	1,305		
dropped out of the survey after being assigned to a condition	0	1,305	0	0
were associated with a duplicate IP Address	6	1,299	4	2
were associated with a duplicate worker ID	0	1,299	0	0
misreported the worker ID	22	1,277	10	12

Final sample: N = 1,277 (626 in the Belief Distribution condition; 651 in the Control condition)

Study 6. Target sample size: N = 1,300
Starting N = 1,581

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Confidence Interval condition	# excluded from Belief Distribution condition	# excluded from Control condition	# excluded from Multiple Guesses condition
dropped out of the survey before being assigned to a condition	287	1,294				
dropped out of the survey after being assigned to a condition	0	1,294				
were associated with a duplicate IP Address	54	1,240	14	14	19	7
were associated with a duplicate worker ID	10	1,230	3	4	1	2
misreported the worker ID	3	1,227	0	1	2	0
failed the attention check	41	1,186	14	8	9	10

Final sample: N = 1,186 (293 in the Confidence Interval condition; 293 in the Belief Distribution condition; 296 in the Control condition; 304 in the Multiple Guesses condition)

Study 7. Target sample size: N = 1,300
Starting N = 1,408

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Confidence Interval condition	# excluded from Belief Distribution condition	# excluded from Control condition	# excluded from Multiple Guesses condition
dropped out of the survey before the attention check	90	1,318				
failed the attention check	18	1,300				
dropped out of the survey after being assigned to a condition	0	1,300	0	0	0	0
were associated with a duplicate IP Address	47	1,253	11	15	10	11
were associated with a duplicate worker ID	6	1,247	1	2	2	1
misreported the worker ID	5	1,242	1	0	4	0

Final sample: N = 1,242 (310 in the Confidence Interval condition; 307 in the Belief Distribution condition; 311 in the Control condition; 314 in the Multiple Guesses condition)

Study 8. Target sample size: N = 1,300
Starting N = 1,330

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Control condition	# excluded from Belief Distribution condition	# excluded from Choosing Possibilities condition	# excluded from Surprise condition
dropped out of the survey before the attention check	19	1,311				
failed the attention check	10	1,301				
dropped out of the survey after being assigned to a condition	0	1,301	0	0	0	0
were associated with a duplicate IP Address	9	1,292	4	2	2	1
were associated with a duplicate worker ID	0	1,292	0	0	0	0
misreported the worker ID	17	1,275	7	1	3	6

Final sample: N = 1,275 (322 in the Control condition; 313 in the Belief Distribution condition; 324 in the Choosing Possibilities condition; 316 in the Surprise condition)

Study 9. Target sample size: N = 1,300
Starting N = 1,561

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Control condition	# excluded from Belief Distribution condition	# excluded from Choosing Possibilities condition	# excluded from Choosing Possibilities + Belief Distribution condition
dropped out of the survey before being assigned to a condition	263	1,298				
dropped out of the survey after being assigned to a condition	0	1,298				
were associated with a duplicate IP Address	80	1,218	20	20	20	20
were associated with a duplicate worker ID	15	1,203	5	3	5	2
misreported the worker ID	11	1,192	3	2	2	4
failed the attention check	87	1,105	22	20	22	23

Final sample: N = 1,105 (272 in the Control condition; 281 in the Belief Distribution condition; 281 in the Choosing Possibilities condition; 271 in the Choosing Possibilities + Belief Distribution condition)

Study 10. Target sample size: N = 1,300
Starting N = 1,465

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Control condition	# excluded from Belief Distribution condition	# excluded from Choosing Possibilities condition	# excluded from Choosing Possibilities + Belief Distribution condition
dropped out of the survey before being assigned to a condition	380	1,085				
dropped out of the survey after being assigned to a condition	0	1,085				
were associated with a duplicate IP Address	51	1,034	12	11	14	14
were associated with a duplicate worker ID	9	1,025	3	0	1	5
misreported the worker ID	4	1,021	0	1	1	2
failed the attention check	75	946	14	22	19	20

Final sample: N = 946 (253 in the Control condition; 236 in the Belief Distribution condition; 242 in the Choosing Possibilities condition; 215 in the Choosing Possibilities + Belief Distribution condition)

Study S1. Target sample size: N = 2,000
Starting N = 2,142

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Belief Distribution condition	# excluded from Control condition
dropped out of the survey before the attention check	129	2,013		
failed the attention check	6	2,007		
dropped out of the survey after being assigned to a condition	0	2,007	0	0
were associated with a duplicate IP Address	75	1,932	26	49
were associated with a duplicate worker ID	7	1,925	3	4
misreported the worker ID	10	1,915	4	6

Final sample: N = 1,915 (959 in the Belief Distribution condition; 956 in the Control condition)

Study S2. Target sample size: N = 200
Starting N = 276

Reason for exclusions	Total # of exclusions	# of remaining participants
dropped out of the survey before the attention check	0	276
failed the attention check	75	201
dropped out of the survey after the attention check	0	201
were associated with a duplicate IP Address	17	184
were associated with a duplicate worker ID	0	184
misreported the worker ID	1	183

Final sample: N = 183

Study S3. Target sample size: N = 1,000
Starting N = 1,079

Reason for exclusions	Total # of exclusions	# of remaining participants	# excluded from Confidence Interval condition	# excluded from Belief Distribution condition	# excluded from Control condition
dropped out of the survey before the attention check	58	1,021			
failed the attention check	18	1,003			
dropped out of the survey after the attention check	0	1,003	0	0	0
were associated with a duplicate IP Address	21	982	7	9	5
were associated with a duplicate worker ID	5	977	1	1	3
misreported the worker ID	3	974	1	2	0

Final sample: N = 974 (318 in the Confidence Interval condition; 319 in the Belief Distribution condition; 337 in the Control condition)

Study S4. Target sample size: N = 200
Starting N = 219

Reason for exclusions	Total # of exclusions	# of remaining participants
dropped out of the survey before the attention check	9	210
failed the attention check	0	210
dropped out of the survey after the attention check	13	197
were associated with a duplicate IP Address	8	189
were associated with a duplicate worker ID	1	188
misreported the worker ID	2	186

Final sample: N = 186

Study S5. Target sample size: N = 400
Starting N = 486

Reason for exclusions	Total # of exclusions	# of remaining participants
failed the attention check	6	480
dropped out of the survey	91	389
were associated with a duplicate IP Address	15	374
were associated with a duplicate worker ID	2	372
misreported the worker ID	3	369

Final sample: N = 369

Notes.

1. For all studies, we requested the target sample size on Prolific/MTurk, so deviations of the number of responses from the target sample size were due to the platform and outside of our control.

2. Studies 9 and 10 (predictions for NBA games) were posted on the survey platform on the morning of the game day (February 12, 2021 and February 19, 2021, respectively). We pre-registered to stop data collection at 8pm Eastern Time (the starting time of the earliest game in our stimuli) if we do not reach our target sample size.
3. In studies with sports predictions (Studies 1, 2, 6, 9, and 10), we asked participants an attention check question at the end of the survey. In Study 2 (where participants made predictions for 2 games), we asked participants to choose the two games they predicted. In Studies 1, 6, 9, and 10 (where participants made predictions for 4 games), we asked participants to choose the game that they did NOT predict. Per our pre-registration, we manually excluded participants who failed this attention check question.
4. In studies with preference/behavior predictions (Studies 3-5, 7-8, S1, and S3), participants first answered the set of preference/behavior questions for themselves. We then asked them to choose the question they did NOT respond to. Participants who failed this attention check question were automatically excluded from the survey.
5. Studies S2, S4, and S5 were not pre-registered. For consistency, we applied the same exclusion rules pre-registered in other studies.

Supplement 2: Study 2 Results with All Participants

In this section, we report the results for the full dataset in Study 2. For the results reported in the paper, we only included participants who met the pre-registered pre-screening criteria (N = 583), i.e., self-reported to have watched an entire NFL game and identified themselves as NFL fans. We present the results including all participants after excluding duplicate responses (N = 812) below.

As in the analyses reported in the main text, we regressed each of the dependent measures on the experimental conditions (dummy-coded), while including fixed effects for the predicted team and clustering standard errors by participant. Relative to providing a 90% confidence interval, giving the entire belief distribution directionally increased confidence (b = .17, SE = .15, t = 1.15, p = .250 for confidence rating; b = 2.84, SE = 1.88, t = 1.51, p = .132 for likelihood estimate; b = .05, SE = .04, t = 1.27, p = .206 for willingness to wager). Relative to providing no belief distribution or confidence interval (Control condition), providing the belief distribution significantly increased participants' confidence in their predictions, b = .63, SE = .15, t = 4.21, p < .001, their likelihood estimates, b = 4.56, SE = 1.88, t = 2.42, p = .016, and directionally increased their likelihood to wager on the predictions, b = .05, SE = .04, t = 1.31, p = .190. Taken together, these results are consistent what we reported in the paper with 583 fans only. We found no evidence that providing a belief distribution reduced confidence in one's predictions.

Supplement 3: Stimuli in Sports Prediction Studies (Studies 1-2, 6, 9-10)

Table S2 presents the set of games included in all sports prediction studies.

Table S2. Studies 1, 2, 6, 9 and 10: Sports Prediction Results by Game

Study	Prediction	Game Date	Game		Actual Points Scored	
			Visiting team	*Home team*	*Visiting team*	*Home team*
1	Winner in the NFL game	October 11, 2020	Eagles	Steelers*	29	38
			Raiders*	Chiefs	40	32
			Colts	Browns*	23	32
			Bills	Titans*	16	42
2	Points scored by a randomly selected team in the NFL game	November 22, 2020	Titans*	Ravens	30	24
			Patriots	Texans*	20	27
			Dolphins	Broncos*	13	20
			Packers	Colts*	31	34
6	Points scored by a randomly selected team in the NFL game	November 1, 2020	Steelers*	Ravens	28	24
			Colts*	Lions	41	21
			Chargers	Broncos*	30	31
			49ers	Seahawks*	27	37
9	Points scored by a randomly selected team in the NBA game	February 12, 2021	Clippers*	Bulls	125	106
			Bucks	Jazz*	115	129
			Thunder	Nuggets*	95	97
			Grizzlies	Lakers*	105	115
10	Points scored by a randomly selected team in the NBA game	February 19, 2021	Pistons	Grizzlies*	95	109
			Suns*	Pelicans	132	114
			Mavericks [1]	Rockets [1]	N/A	N/A
			Jazz	Clippers*	112	116

Note. Within each row, the winner of the game is marked with an asterick.

[1] The Mavericks vs. Rockets game (on Feb 19, 2021) was cancelled after we posted our study.

Supplement 4: Evidence that Likelihood Estimates are Overconfident

Our finding that the belief distribution elicitation tends to increase confidence has different implications depending on whether participants were *under*confident or *over*confident at the baseline. If participants were initially underconfident in the accuracy of their predictions, then constructing a belief distribution improved their calibration. On the other hand, if they were already overconfident at the baseline, then eliciting a belief distribution exacerbated this tendency.

Past literature on overconfidence typically used hit rate across several predictions as a measure of calibration. In our studies, participants made up to four predictions, and so we did not have enough observations per participant for a good measure of individual calibration.

Instead, we looked at aggregate-level calibration. For each prediction item, participants responded to a likelihood estimate question "In your opinion, how likely is your prediction to be within [X] of the correct answer?" [X] was different across studies (see Table 1 in the paper for exact wordings). For each prediction item, we derived the percentage of observations whose best estimate responses fell within the [X] range of the correct answer and used this as a proxy for aggregate-level accuracy. The only exception is Study 1, where participants made a binary prediction and answered the question "In your opinion, how likely are [your predicted winner] to win?" For the games in Study 1, we derived instead the percentage of observations that correctly predicted the winner. The results for this measure are reported in the "Accuracy" rows in Table S3.1.

We compared this measure to participants' average response to the likelihood estimate, reported in the "Estimate" row in Table S3.1. If participants were underconfident, then the numbers in the "Accuracy" row should be greater than those in the "Estimate" row. However, if the "Estimate" results are greater than the "Accuracy" results, then participants were overconfident.

As shown in Table S3.1, participants overestimated the percentage of times their answers would be accurate by a considerable amount in all conditions in all studies. In other words, even at the baseline, participants were vastly overconfident in their predictions. Constructing a belief distribution further magnified this tendency.

Table S3.1. Comparisons Between Participants' Prediction Accuracy and Likelihood Estimate Responses

Study	Domain	*Measure*	Conditions				
1	NFL		-	Belief Distribution	Confidence Interval		
		Accuracy	-	58.94%	60.11%		
		Estimate	-	68.46%	69.37%		
2	NFL		Control	Belief Distribution	Confidence Interval		
		Accuracy	38.60%	36.10%	38.42%		
		Estimate	58.72%	63.04%	59.92%		
3	Preference		Control	Belief Distribution / Best Estimate First	Belief Distribution / Best Estimate Last	Confidence Interval / Best Estimate First	Confidence Interval / Best Estimate Last
		Accuracy	22.78%	23.29%	22.52%	25.68%	22.01%
		Estimate	47.13%	51.31%	51.77%	48.78%	47.56%
4	Preference		Control / Low Precision	Belief Distribution / Low Precision	Control / High Precision	Belief Distribution / High Precision	
		Accuracy	16.67%	16.67%	15.57%	17.39%	
		Estimate	51.10%	51.18%	56.18%	54.21%	
5	Preference		Control / Moderate	Belief Distribution / Moderate	Control / Extreme	Belief Distribution / Extreme	
		Accuracy	26.51%	29.04%	58.52%	58.74%	
		Estimate	50.47%	54.46%	73.76%	77.67%	
6	NFL		Control	Belief Distribution	Confidence Interval	Multiple Guesses	
		Accuracy	25.17%	26.88%	23.46%	27.55%	
		Estimate	49.28%	50.95%	47.35%	49.43%	
7	Preference		Control	Belief Distribution	Confidence Interval	Multiple Guesses	
		Accuracy	26.05%	25.81%	26.13%	25.72%	
		Estimate	47.23%	50.03%	48.89%	51.24%	
8	Preference		Control	Belief Distribution	Choosing Possibilities	Surprise	
		Accuracy	16.93%	18.93%	17.98%	20.89%	
		Estimate	49.77%	52.10%	47.96%	52.61%	
9	NBA		Control	Belief Distribution	Choosing Possibilities	Choosing Possibilities + Belief Distribution	
		Accuracy	33.27%	35.32%	33.99%	33.76%	
		Estimate	56.59%	62.04%	59.25%	61.84%	
10	NBA		Control	Belief Distribution	Choosing Possibilities	Choosing Possibilities + Belief Distribution	
		Accuracy	41.63%	41.95%	43.53%	40.78%	
		Estimate	55.79%	61.00%	57.44%	60.50%	

Notes.

1. In each study, participants responded to a likelihood estimate question in the form of "In your opinion, how likely is your prediction to be within [X] of the correct answer?" The ***"Accuracy"*** rows in this table capture the percentage of all participants' responses that fall within [X] of the correct answer. The only exception is Study 1, where participants made a binary prediction without giving a best estimate prediction. We therefore calculated the percentage of responses that predicted the correct winner instead for Study 1.
2. The ***"Estimate"*** rows in this table show participants' average response to the likelihood estimate question.

Furthermore, we correlated these two measures within each condition to see if participants' calibration differed by condition. As Table S3.2 below shows, this correlation does not differ meaningfully across conditions, suggesting that the manipulations did not change how calibrated participants were but merely shifted the confidence level.

Table S3.2. Correlation Between Participants' Prediction Accuracy and Likelihood Estimate Responses

Study	Conditions					
1	**All conditions**	**-**	**Belief Distribution**	**Confidence Interval**		
	r = -.12 (p < .001)		r = -.13 (p < .001)	r = -.12 (p < .001)		
2	**All conditions**	**Control**	**Belief Distribution**	**Confidence Interval**		
	r = .01 (p = .741)	r = .02 (p = .767)	r = -.01 (p = .778)	r = .03 (p = .526)		
3	**All conditions**	**Control**	**Belief Distribution / Best Estimate First**	**Belief Distribution / Best Estimate Last**	**Confidence Interval / Best Estimate First**	**Confidence Interval / Best Estimate Last**
	r = -.02 (p = .129)	r = -.03 (p = .308)	r = -.02 (p = .347)	r = -.02 (p = .548)	r = -.01 (p = .574)	r = -.01 (p = .780)
4	**All conditions**	**Control / Low Precision**	**Belief Distribution / Low Precision**	**Control / High Precision**	**Belief Distribution / High Precision**	
	r = -.04 (p = .003)	r = -.05 (p = .086)	r = -.06 (p = .033)	r = -.03 (p = .287)	r = -.03 (p = .318)	
5	**All conditions**	**Control / Moderate**	**Belief Distribution / Moderate**	**Control / Extreme**	**Belief Distribution / Extreme**	
	r = .34 (p < .001)	r = .03 (p = .396)	r = .04 (p = .227)	r = .42 (p < .001)	r = .41 (p < .001)	
6	**All conditions**	**Control**	**Belief Distribution**	**Confidence Interval**	**Multiple Guesses**	
	r = -.002 (p = .867)	r = .01 (p = .736)	r = -.02 (p = .463)	r = -.02 (p = .560)	r = .01 (p = .646)	
7	**All conditions**	**Control**	**Belief Distribution**	**Confidence Interval**	**Multiple Guesses**	
	r = -.02 (p = .265)	r = -.02 (p = .446)	r = -.01 (p = .661)	r = -.03 (p = .359)	r = -.003 (p = .931)	
8	**All conditions**	**Control**	**Belief Distribution**	**Choosing Possibilities**	**Surprise**	
	r = -.03 (p = .054)	r = -.01 (p = .679)	r = -.06 (p = .051)	r = .02 (p = .561)	r = -.07 (p = .019)	
9	**All conditions**	**Control**	**Belief Distribution**	**Choosing Possibilities**	**Choosing Possibilities + Belief Distribution**	
	r = .02 (p = .112)	r = .02 (p = .598)	r = .002 (p = .952)	r = .04 (p = .211)	r = .04 (p = .235)	
10	**All conditions**	**Control**	**Belief Distribution**	**Choosing Possibilities**	**Choosing Possibilities + Belief Distribution**	
	r = .07 (p < .001)	r = .08 (p = .028)	r = .07 (p = .057)	r = .08 (p = .025)	r = .04 (p = .351)	

Note. This table reports the correlation between participants' prediction accuracy and participants' likelihood estimate responses within each condition (derived the same way as described in Table S3.1 note).

Supplement 5: Confidence Interval Width Results (Studies 1-3, 6, 7, S3)

In this section, we present the 90% confidence interval width results for the Belief Distribution condition and the Confidence Interval condition. Past research found that the confidence intervals derived from the full belief distributions are wider than those directly stated by participants. Although this is not the primary aim of our paper, we report the results of this measure to conceptually replicate past findings.

In the Belief Distribution condition, we used the algorithm developed by Haran et al (2010) to derive the 90% confidence interval for each elicited distribution. The algorithm requires the range of outcomes to be bounded by a minimum and a maximum. In studies with sports predictions (Studies 1, 2, and 6), the ranges in the belief distribution interface were unbounded (e.g., the lowest category was "lose by more than 30 points" and the highest category was "win by more than 30 points" in Study 1). Therefore, we determined and pre-registered the minimum and the maximum based on the range partition in each study: In Study 1, we used -40 and 40; in Study 2, we used 0 and 48; in Study 6, we used 0 and 40. In studies with preference predictions (Studies 3, 7, and S3), we used the natural boundaries of 0% and 100%.

To keep the upper bound and lower bound consistent, we pre-registered to winsorize the values in the Confidence Interval condition to the same minimum and maximum. For example, in Study 1, we winsorized any confidence interval values of less than -40 to -40 and any values greater than 40 to 40. In addition, for "backwards" confidence intervals – ones with higher lower bounds than upper bounds – we pre-registered to treat them as equal to 0 in Study 1 and as equal to the absolute difference between the two values in all other studies. Results excluding responses with "backwards" confidence intervals do not differ meaningfully.

Table S4 presents all comparisons of the width of the 90% confidence interval between the Belief Distribution condition and the Confidence Interval condition. Consistent with past findings, in most cases, the confidence intervals derived from the Belief Distribution condition were significantly wider than those stated in the Confidence Interval condition. There were a few exceptions where the intervals did not differ across the two conditions (in Study 2), or the intervals were even significantly wider in the Confidence Interval condition (in Study 6). This may have occurred due to the way the ranges were bounded in the Belief Distribution condition. That is, participants' responses in the Belief Distribution condition were constrained by the range of outcomes provided. For example, in Study 2, the lowest category and the highest category were "score 6 points or fewer" and "score 43 points or more," and we set the minimum and the maximum to be 0 and 48. Thus, participants' belief distributions cannot exceed those boundaries in the Belief Distribution condition. But those in the Confidence Interval condition didn't have such constraints and thus produced wider confidence intervals.

Additional confidence interval width results for the Belief Distribution condition

In Study 4, we pre-registered to compare the confidence interval width imputed from the belief distribution between the Low Precision condition and the High Precision condition. We are not reporting this result because we realize in subsequent analysis that the confidence interval width is on different scales in these two conditions.

In Study 5, we pre-registered as secondary analyses to regress the confidence interval width imputed from the belief distribution on the Extreme Answer condition, including fixed effects for predicted item and clustering standard errors by participant. We found that the 90% confidence interval derived from the belief distribution was significantly narrower for extreme questions (*M*

= 22.44, *SD* = 21.20) than for moderate questions (*M* = 39.35, *SD* = 22.09), *b* = -16.91, clustered *SE* = 1.05, *t* = -16.06, *p* < .001.

Table S4. Confidence Interval Width Results by Game/Question

			Belief Distribution condition		Confidence Interval condition		
Study	Prediction	Game/Item	Mean	SD	Mean	SD	Pairwise comparison
1[p]	Point differential in the NFL game	Eagles vs. Steelers	**38.27**	14.47	16.26	15.52	$t(509) = 16.56$, $p < .001$
		Raiders vs. Chiefs	**37.45**	14.44	17.53	16.84	$t(509) = 14.32$, $p < .001$
		Colts vs. Browns	**39.99**	16.73	21.38	19.07	$t(509) = 11.70$, $p < .001$
		Bills vs. Titans	**39.97**	16.83	20.30	18.51	$t(509) = 12.55$, $p < .001$
		All Games	$b = 20.05$, clustered $SE = 1.27$, $p < .001$				
2	Points scored by a randomly selected team in the NFL game	Titans vs. Ravens	21.73	7.12	24.29	11.98	$t(205) = -1.84$, $p = .067$
		Patriots vs. Texans	21.90	7.81	20.77	11.06	$t(194) = .82$, $p = .415$
		Dolphins vs. Broncos	21.65	7.86	21.18	11.09	$t(190) = .33$, $p = .738$
		Packers vs. Colts	22.89	7.67	22.79	10.72	$t(183) = .07$, $p = .946$
		All Games	$b = -.25$, clustered $SE = .92$, $p = .784$				
3[p,1]	Percentage of all survey respondents who have a particular preference	Thanksgiving vs. Christmas	**39.31**	23.53	35.17	25.24	$t(1,453) = 3.23$, $p = .001$
		See the future vs. Change the past	**42.69**	24.01	37.18	25.45	$t(1,453) = 4.25$, $p < .001$
		1 wish today vs. 3 wishes in 5 years	**42.18**	23.55	35.82	25.76	$t(1,453) = 4.91$, $p < .001$
		More money vs. More time	40.96	23.97	38.42	26.03	$t(1,453) = 1.94$, $p = .053$
		All Items	$b = 4.64$, clustered $SE = 1.08$, $p < .001$				
6	Points scored by a randomly selected team in the NFL game	Steelers vs. Ravens	19.44	7.17	**22.73**	8.50	$t(584) = -5.07$, $p < .001$
		Colts vs. Lions	19.25	7.39	**22.66**	9.16	$t(584) = -4.96$, $p < .001$
		Chargers vs. Broncos	19.59	7.10	**21.95**	8.62	$t(584) = -3.63$, $p < .001$
		49ers vs. Seahawks	19.19	7.07	**22.91**	8.67	$t(584) = -5.69$, $p < .001$
		All Games	$b = -3.20$, clustered $SE = .58$, $p < .001$				
7	Percentage of all survey respondents who have a particular preference	Pancakes vs. Waffles	**41.48**	22.36	33.18	24.42	$t(615) = 4.40$, $p < .001$
		Invisibility vs. Time travel	**43.59**	23.73	36.14	25.05	$t(615) = 3.79$, $p < .001$
		Music vs. Podcast	**41.53**	23.11	34.06	24.12	$t(615) = 3.93$, $p < .001$
		Coffee smell vs. Cookies smell	**42.43**	23.02	34.89	24.66	$t(615) = 3.93$, $p < .001$
		All Items	$b = 7.69$, clustered $SE = 1.68$, $p < .001$				
S3[p]	Percentage of all survey respondents who have a particular preference	Exercise vs. Reading	**40.61**	22.55	33.60	24.11	$t(635) = 3.79$, $p < .001$
		Blog Post (Yes vs. No)	**39.11**	21.92	33.08	25.13	$t(635) = 3.23$, $p = .001$
		Instagram (Yes vs. No)	35.50	22.21	38.92	26.14	$t(635) = -1.78$, $p = .075$
		Fall vs. Winter	38.31	23.03	35.58	24.58	$t(635) = 1.45$, $p = .149$
		All Items	$b = 3.08$, clustered $SE = 1.57$, $p = .049$				

Notes. This table presents all comparisons of the confidence interval width results between the Belief Distribution condition and the Confidence Interval condition. Studies without the Confidence Interval condition (Studies 4-5, 8-10) are necessarily omitted in this table.

Within each row, **boldface** indicates that the confidence interval is significantly wider in this condition.

Within each study, the regression includes fixed effects for game (Study 1) / team (Studies 2 and 6) / item (Studies 3, 7, S3) and clusters standard errors by participant.

[p] indicates that the confidence interval width was pre-registered as a primary DV.

[1] The Best Estimate First and the Best Estimate Second conditions don't differ significantly from each other, $t(5,818) = .17$, $p = .862$, and are collapsed for this study.

Supplement 6: Allocation Results in the Belief Distribution Condition

In the paper, we suggest that the process of allocating probabilities to different outcomes might inadvertently reinforce people's initial beliefs. This mechanism would predict that the more mass participants allocate to categories surrounding their initial belief, the more confident they should be in their predictions. Table S5 reports results consistent with this prediction.

Table S5 has two sections. In the section to the left, titled **"Average Percentage of Mass Allocated to…"**, we report the average amount of mass participants allocated to (1) the bin containing their best estimate, (2) the three bins surrounding their best estimate, and (3) the half of the range corresponding to the opposite prediction. If participants constructed belief distributions in a way that confirmed their best estimate, we would expect our confidence measures to be positively correlated with (1) and (2) and be negatively correlated with (3). That is indeed what we found.

In the section to the right, we report the percentage of participants allocating (1) over 50% of the mass, (2) over 75% of the mass, and (3) over 90% of the mass to the three bins surrounding the best estimate. The results show that the vast majority of participants allocated over half of the mass to those three bins. Many of them allocated even more than 75% and 90% of the mass to bins surrounding the best estimate. This suggests that most participants gave belief distributions with the greatest density around their initial predictions and thus provides corroborating evidence to our proposed mechanism.

Table S5. Allocation Results in the Belief Distribution Condition

Study	Domain	Average Percentage of Mass Allocated to …			Percentage of Participants Allocating		
					> 50% of Mass	> 75% of Mass	> 90% of Mass
		the Bin Containing the Best Estimate	the Three Bins Surrounding the Best Estimate	the Half of the Range Corresponding to the Opposite Prediction	to the Three Bins Surrounding the Best Estimate		
2	**NFL**	40.65%	75.40%	-			
Correlation with Confidence Rating		r = .28 (p < .001)	r = .26 (p < .001)		95.25%	77.09%	59.13%
Correlation with Likelihood Estimate		r = .25 (p < .001)	r = .27 (p < .001)				
Correlation with Wager Likelihood		r = .12 (p = .006)	r = .13 (p = .003)				
3 (Best Estimate First)	**Preference**	35.33%	66.35%	50.19%			
Correlation with Confidence Rating		r = .21 (p < .001)	r = .11 (p < .001)	r = -.12 (p < .001)	71.78%	42.60%	25.75%
Correlation with Likelihood Estimate		r = .28 (p < .001)	r = .22 (p < .001)	r = -.18 (p < .001)			
3 (Best Estimate Last)	**Preference**	30.94%	64.35%	47.78%			
Correlation with Confidence Rating		r = .16 (p < .001)	r = .09 (p < .001)	r = -.15 (p < .001)	66.80%	41.80%	25.55%
Correlation with Likelihood Estimate		r = .20 (p < .001)	r = .17 (p < .001)	r = -.17 (p < .001)			
4 (High Precision)	**Preference**	32.75%	59.13%	42.40%			
Correlation with Confidence Rating		r = -0.01 (p = .612)	r = -.20 (p < .001)	r = -.06 (p = .046)	62.54%	32.61%	14.21%
Correlation with Likelihood Estimate		r = .18 (p < .001)	r = .09 (p = .002)	r = -.16 (p < .001)			
4 (Low Precision)	**Preference**	28.98%	60.40%	52.59%			
Correlation with Confidence Rating		r = .02 (p = .599)	r = -.09 (p = .004)	r = -.21 (p < .001)	63.95%	33.42%	18.54%
Correlation with Likelihood Estimate		r = .17 (p < .001)	r = .08 (p = .008)	r = -.25 (p < .001)			
5 (Extreme)	**Preference**	60.33%	77.82%	45.27%			
Correlation with Confidence Rating		r = .43 (p < .001)	r = .31 (p < .001)	r = -.20 (p < .001)	78.64%	72.25%	62.72%
Correlation with Likelihood Estimate		r = .41 (p < .001)	r = .33 (p < .001)	r = -.19 (p < .001)			
5 (Moderate)	**Preference**	33.93%	65.30%	53.43%			
Correlation with Confidence Rating		r = .16 (p < .001)	r = .07 (p = .026)	r = -.09 (p = .008)	72.05%	42.69%	24.70%
Correlation with Likelihood Estimate		r = .19 (p < .001)	r = .15 (p < .001)	r = -.14 (p < .001)			
6	**NFL**	37.14%	71.11%	-			
Correlation with Confidence Rating		r = .21 (p < .001)	r = .15 (p < .001)		86.52%	41.21%	15.87%
Correlation with Likelihood Estimate		r = .27 (p < .001)	r = .20 (p < .001)				
7	**Preference**	32.03%	63.47%	52.45%			
Correlation with Confidence Rating		r = .12 (p < .001)	r = .02 (p = .552)	r = -.15 (p < .001)	68.08%	39.50%	22.64%
Correlation with Likelihood Estimate		r = .23 (p < .001)	r = .17 (p < .001)	r = -.17 (p < .001)			
8	**Preference**	37.02%	69.27%	52.62%			
Correlation with Confidence Rating		r = .21 (p < .001)	r = .12 (p < .001)	r = -.23 (p < .001)	74.84%	47.44%	28.91%
Correlation with Likelihood Estimate		r = .28 (p < .001)	r = .23 (p < .001)	r = -.21 (p < .001)			
9	**NBA**	47.45%	83.41%	-			
Correlation with Confidence Rating		r = .22 (p < .001)	r = .15 (p < .001)		93.59%	71.44%	45.02%
Correlation with Likelihood Estimate		r = .30 (p < .001)	r = .27 (p < .001)				
10	**NBA**	43.13%	82.08%	-			
Correlation with Confidence Rating		r = .21 (p < .001)	r = .14 (p < .001)		91.10%	70.87%	45.76%
Correlation with Likelihood Estimate		r = .28 (p < .001)	r = .24 (p < .001)				

Supplement 7: Effect Sizes by Presentation Order

In all our studies, participants made multiple predictions. Does the effect size (the extent to which giving the belief distribution increases reported confidence) change over time? In Table S6, we report the effect size estimates (Cohen's *d*s comparing the Belief Distribution condition and the Control condition) by the order in which participants made the prediction. In most studies, the effect sizes for both the confidence rating question and the likelihood estimation question declined over the course of the study.

Table S6. Effect Size Estimate (Cohen's *d*s) by the Order of Presented Items

		Confidence Rating							
Study	**Domain**	**1st presented item**		**2nd presented item**		**3rd presented item**		**4th presented item**	
		d	95% CI	*d*	95% CI	*d*	95% CI	*d*	95% CI
2	**NFL**	0.42	[0.22, 0.63]	0.33	[0.12, 0.53]	-		-	
3	**Preference**	0.32	[0.19, 0.44]	0.07	[-0.06, 0.19]	0.12	[-0.01, 0.25]	0.09	[-0.04, 0.21]
4		0.16	[0.05, 0.27]	-0.04	[-0.15, 0.07]	0.01	[-0.10, 0.12]	-0.05	[-0.16, 0.07]
5		0.23	[0.12, 0.34]	0.16	[0.05, 0.27]	0.15	[0.04, 0.26]	-	
6	**NFL**	0.36	[0.20, 0.52]	0.29	[0.12, 0.45]	0.31	[0.15, 0.47]	0.24	[0.07, 0.40]
7	**Preference**	0.27	[0.11, 0.43]	0.04	[-0.12, 0.20]	0.08	[-0.08, 0.24]	0.10	[-0.05, 0.26]
8		0.29	[0.14, 0.45]	0.11	[-0.04, 0.27]	0.18	[0.02, 0.33]	-0.02	[-0.17, 0.14]
9	**NBA**	0.32	[0.15, 0.48]	0.25	[0.08, 0.41]	0.22	[0.05, 0.39]	0.21	[0.04, 0.37]
10		0.31	[0.13, 0.48]	0.20	[0.03, 0.38]	0.32	[0.14, 0.50]	0.20	[0.02, 0.37]
		Likelihood Estimate							
Study	**Domain**	**1st presented item**		**2nd presented item**		**3rd presented item**		**4th presented item**	
		d	95% CI	*d*	95% CI	*d*	95% CI	*d*	95% CI
2	**NFL**	0.22	[0.02, 0.42]	0.18	[-0.03, 0.38]	-		-	
3	**Preference**	0.35	[0.22, 0.48]	0.12	[-0.01, 0.24]	0.10	[-0.03, 0.23]	0.09	[-0.04, 0.21]
4		0.05	[-0.06, 0.16]	-0.05	[-0.16, 0.06]	-0.05	[-0.16, 0.06]	-0.08	[-0.20, 0.03]
5		0.15	[0.04, 0.26]	0.13	[0.02, 0.24]	0.17	[0.06, 0.28]	-	
6	**NFL**	0.08	[-0.08, 0.24]	0.11	[-0.05, 0.27]	0.04	[-0.13, 0.20]	0.06	[-0.10, 0.22]
7	**Preference**	0.19	[0.03, 0.35]	0.03	[-0.12, 0.19]	0.10	[-0.06, 0.26]	0.11	[-0.05, 0.26]
8		0.23	[0.07, 0.38]	0.03	[-0.13, 0.18]	0.12	[-0.04, 0.27]	-0.02	[-0.18, 0.13]
9	**NBA**	0.25	[0.09, 0.42]	0.25	[0.08, 0.42]	0.25	[0.08, 0.41]	0.16	[-0.01, 0.33]
10		0.17	[-0.01, 0.34]	0.23	[0.05, 0.41]	0.25	[0.08, 0.43]	0.16	[-0.02, 0.34]

Note. The Cohen's *d*s compare the Belief Distribution condition and the Control condition (Belief Distribution condition minus the Control condition). A positive sign reflects that the Belief Distribution condition increased confidence compared to the Control condition; a negative sign reflects that the Belief Distribution condition reduced confidence compared to the Control condition.

Supplement 8: Study S1

In Footnote 2 in the Introduction, we mentioned Study 3 in Haran et al (2010), where they documented that providing a belief distribution for previous estimates increased the width of confidence intervals constructed for subsequent estimates. In that study, participants were asked to estimate the year in which 16 U.S. presidents were first elected. They were randomly assigned to one of two conditions in a 2 (Elicitation Method: Belief Distribution vs. Confidence Interval) x 2 (Elicitation Order: Belief Distribution First vs. Belief Distribution Last) mixed design with the Elicitation Method as a within-subjects factor and the Elicitation Order as a between-subjects factor. That is, all participants gave estimates for 8 presidents with a belief distribution and gave estimates for the other 8 presidents with a 90% confidence interval. Participants assigned to the Belief Distribution First condition provided belief distributions for the first 8 items and provided confidence intervals for the next 8 items. The order was reversed for those assigned to the Belief Distribution Last condition. The authors replicated their main finding that the 90% confidence intervals derived from the belief distribution were significantly wider than those elicited in the Confidence Interval condition. Moreover, they found that within the Confidence Interval conditions, the intervals were significantly wider for the last eight items than for the first eight items, $t(332) = 3.25$, $p = .001$. The authors interpreted these results as evidence that giving a belief distribution "had a carryover effect on subsequent confidence interval estimates, leading judges to consider a wider range of values in their estimates."

However, we are worried that the design of that study potentially confounded the elicitation method with the order. For example, it is possible that participants became less confident in how much they knew about the election years of U.S. presidents as they proceeded through the task and that they therefore provided wider confidence intervals for later items than for early items. As a

result, the wider confidence intervals for the last 8 items (after completing the belief distribution task) could be caused by the belief distribution task (as suggested by the authors) or by item order. Haran et al (2010) tried to rule out this possibility in an analysis presented in their Figure 4, but that analysis is necessarily inconclusive. A cleaner test is needed.

To cleanly test whether giving a belief distribution increases the width of subsequent confidence intervals, we conducted Study S1. We addressed the abovementioned confound by asking participants to provide a confidence interval after giving the belief distribution for the same prediction item. Study S1 followed a similar procedure as in other studies in our paper. Participants made predictions of other participants' responses to four preference and behavior questions. For each item, they first provided a best estimate. Half of the participants then gave a belief distribution around the best estimate. Finally, instead of reporting their confidence using rating scales, all participants gave a 90% confidence interval around their best estimate. If Haran et al (2010)'s findings were driven by the influence of the belief distribution task rather than a decrease in confidence over time, then participants who gave their belief distributions in our study should provide wider confidence intervals than those who didn't. However, contrary to Haran et al (2010), we found no carryover effect of the belief distribution task on the width of subsequent confidence intervals.

We do not include this study in the main text because we do not think eliciting confidence intervals this way is a valid assessment of overconfidence (e.g., Langnickel & Zeisberger, 2016). Nevertheless, we wanted to see if we could replicate the carryover effect with our paradigm, which we believe provides a cleaner test of the question.

Method

Participants. We conducted Study S1 using U.S. participants from Prolific. We decided in advance to recruit 2,000 participants. Only participants who passed the attention check at the beginning of the survey were allowed to proceed. We excluded participants whose Prolific IDs or IP addresses appeared more than once (82 exclusions) and excluded participants who misreported Prolific IDs (10 exclusions). This left us with a final sample of 1,915, averaged 37.5 years of age and was 49.2% female.

Procedure. Participants were randomly assigned to one of four conditions in a 2 (Control vs. Belief Distribution) x 2 (Confidence Interval Elicitation: Two-point vs. Range) between-subjects design. Study S1 followed a similar procedure as in other preference prediction studies in the paper, except for one change in the dependent measure. For each prediction item, all participants were asked to provide a confidence interval such that they were 90% sure that the correct answer falls within the interval. We manipulated how the 90% confidence interval was elicited. In the Two-point condition, participants indicated the 5^{th} and the 95^{th} fractiles separately (Soll & Klayman, 2004). Specifically, they were asked to fill out the questions: "I am 95% sure that at least ___ % of participants [would choose the option]" and "I am 95% sure that at most ___% of participants [would choose the option]." In the Range condition, participants indicated the upper and lower bound of the confidence interval together. Specifically, they were asked to answer the question: "I am 90% sure that between ___% and ___% of participants [would choose the option]."

Results and Discussion. Our dependent measure was the width of the 90% confidence interval. This was calculated by subtracting the lower bound from the upper bound. As pre-registered, if participants indicated a "backward" interval – that is, provided a lower bound that is higher than the upper bound – we reversed their responses and subtracted the upper bound from the lower bound.

We pre-registered to use OLS to regress the width of the 90% confidence interval on (1) the Belief Distribution/Control condition (contrast-coded), (2) the Confidence Interval Elicitation condition (contrast-coded), and (3) their interaction. We included the fixed effects for prediction items and clustered standard errors by participant.

Contrary to Haran et al (2010)'s findings, providing a belief distribution (M = 29.33, SD = 18.68) did not increase the width of the confidence interval compared to the Control condition (M = 29.54, SD = 18.74), b = -.21, SE = .74, t = -.29, p = .775. This result held regardless of whether participants specified the upper bound and the lower bound of the confidence interval separately (Two-point/Belief Distribution condition: M = 29.54, SD = 19.66; Two-point/Control condition: M = 29.11, SD = 18.97), b = .44, SE = 1.07, t = .41, p = .683, or entered the two values in one question (Range/Belief Distribution condition: M = 29.12, SD = 17.63; Range/Control condition: M = 29.98, SD = 18.49), b = -.86, SE = 1.03, t = -.84, p = .403. The method of eliciting the confidence interval did not influence the width of the confidence interval (b = -.23, SE = .74, t = -.30, p = .761) and there was no interaction between the two factors (b = 1.30, SE = 1.49, t = .88, p = .381). In sum, first providing a belief distribution did not widen the confidence interval participants subsequently constructed for the same prediction item.

In our exploratory analyses, we derived the 90% confidence intervals from participants' belief distributions. As in most other studies (see Supplement 4), the confidence intervals imputed from the belief distribution (M = 39.74, SD = 20.73) were significantly wider than those directly stated (M = 29.33, SD = 18.68), $t(3,835) = 25.35$, $p < .001$. In fact, 61.4% of the observations in the Belief Distribution condition had a wider confidence interval in the belief distribution than directly stated. In other words, even though participants gave belief distributions that corresponded to wider

confidence intervals, they did not subsequently construct wider confidence intervals as a result of that.

Supplement 9: Study S2

Study S2 was a pre-test for the stimuli used in Study 5. In Study 5, we aimed to provide participants with a similar set of preference / behavior questions but manipulate the extremity of the answers. We came up with six pairs of questions, each with a Moderate version and an Extreme version. We expect that the true percentage of answers would be between 30% and 70% for the Moderate version, and above 90% or below 10% for the Extreme version. The purpose of Study S2 was to pre-test these pairs of questions.

Method

Participants. We conducted Study S2 using U.S. participants from Amazon Mechanical Turk (MTurk). We decided in advance to recruit 200 participants. This study was a pre-test for Study 5 and was therefore not pre-registered. For consistency, we followed the pre-registered exclusion rules for Study 5: Only participants who passed the attention check at the beginning of the survey were allowed to proceed. In addition, we excluded participants whose MTurk IDs or IP addresses appeared more than once (17 exclusions) and participants who misreported MTurk IDs (1 exclusion). This left us with a final sample of 183, averaged 37.5 years of age and was 40.4% female.

Procedure. All participants followed the same study procedure. They answered six preference or behavior questions about themselves at the beginning of the survey (presented in Table S7). For each question, they were randomly assigned to see either the Extreme or the Moderate version. Then, they estimated for each question the percentage of participants choosing a particular option. On the next page, all participants provided the belief distribution around the prediction. We didn't collect any of the confidence measures in this pre-test.

Table S7. Study S2: Wording and True Percentages for Preference/Behavior Questions

Preference/Behavior questions	Category	% of participants who chose the first option
Have you ever visited Nigeria? (Yes/No)	*Extreme*	32%
Have you ever visited France? (Yes/No)	*Moderate*	43%
Do you prefer milk chocolate or wasabi-flavored chocolate?	*Extreme*	96%
Do you prefer milk chocolate or dark chocolate?	*Moderate*	63%
Which ice cream flavor do you prefer: chocolate or cheese?	*Extreme*	97%
Which ice cream flavor do you prefer: chocolate or vanilla?	*Moderate*	54%
On a typical Saturday, do you wake up before 5am?	*Extreme*	31%
On a typical Saturday, do you wake up before 8am?	*Moderate*	59%
Did you read more than 20 books in 2020? (Yes/No)	*Extreme*	47%
Did you read more than 2 books in 2020? (Yes/No)	*Moderate*	80%
Do you have a TV? (Yes/No)	*Extreme*	97%
Do you have an iPad? (Yes/No)	*Moderate*	59%

Results and Discussion

The goal of this study was to select the set of questions to be used in Study 5. As shown in Table S7, three questions met our criteria of the Extreme vs. Moderate distinction: the "chocolate" question, the "ice cream flavor" question, and the "TV/iPad" question. The Extreme version of all these questions had a choice share over 90% or below 10%, while their Moderate version had a choice share between 30% and 70%. Therefore, we used these three questions in Study 5.

Supplement 10: Study S3

Study S3 was a replication of Study 2 in a different prediction domain. We are not including this study in the main text because we think the task was confusing for participants. Specifically, we think that our key measures (e.g., "How confident are you in your prediction that more than half of participants [picked the option you predicted to be the majority option]?" and "How likely is it that more than half of survey respondents [picked the option you predicted to be the majority option]?") were confusing for participants. When we made it less confusing in subsequent studies, we found different results. Nevertheless, we are reporting the study here for completeness.

Method

Participants. We conducted Study S3 using U.S. participants from Prolific. We decided in advance to recruit 1,000 participants. Only participants who passed the attention check at the beginning of the survey were allowed to proceed. We pre-registered to keep the first response only from Prolific IDs or IP addresses that appeared more than once (26 exclusions) and exclude participants who misreported Prolific IDs (3 exclusions). This left us with a final sample of 974, averaged 33.0 years of age, and was 49.3% female.

Procedure. Study S3 followed a similar procedure as Study 2, except for two changes. First, the prediction domain was different. Instead of predicting the outcomes of NFL games as in Study 2, participants in Study S3 predicted other participants' preferences and behaviors. The questions are presented in Table S8.

Table S8. Study S3: Wording and True Percentages for Preference/Behavior Questions

Preference/Behavior questions	% of participants who chose the first option
Do you prefer exercising or reading?	40%
Have you ever written a blog post? (Yes/No)	43%
Do you have an Instagram account? (Yes/No)	77%
Do you prefer the fall season or the winter season?	78%

Second, due to the change of domain, the prediction questions and the wording of the confidence measures were different. For each preference/behavior question, we asked participants to predict which option more survey respondents would choose. Note that this was a binary prediction question and was different from the percentage prediction question we asked in other studies (see Table 2). Participants responded to three questions that served as our dependent measures:

(1) "How confident are you that more survey respondents [picked the option you predicted to be the majority option]?" (1 = Not at all confident, 9 = Extremely confident)

(2) "In your opinion, how likely is it that more survey respondents [picked the option you predicted to be the majority option]?" (possible answers range from 0 % to 100%)

(3) Participants received an additional bonus of 10 cents and could wager any of the 10 cents on their prediction. The amount they wagered would double if their prediction were correct. They responded to the question "How much would you like to wager on your prediction?" (possible answers range from 0 cents to 10 cents)

Results and Discussion

We regressed participants' confidence on the experimental conditions, while including fixed effects for the predicted item and clustering standard errors by participant. We present the results in Figure S1. Compared to providing a 90% confidence interval, giving the entire belief

distribution significantly increased participants' confidence in their predictions (b = .52, SE = .12, t = 4.41, p < .001), directionally increased their likelihood estimates (b = 2.09, SE = 1.18, t = 1.77, p = .077), and had no influence on willingness to wager (b = -.04, SE = .22, t = -.19, p = .851). Compared to only providing a best estimate (the Control condition), giving the entire belief distribution had no influence on the likelihood estimate (b = 1.53, SE = 1.10, t = 1.39, p = .164) and the willingness to wager (b = -.32, SE = .20, t = -1.57, p = .118), but significantly reduced participants' confidence in their predictions (b = -.22, SE = .10, t = -2.20, p = .028). The comparisons with the Control condition on the confidence measure are in the opposite direction of what we consistently found in other studies. Several participants left comments at the end of the survey indicating that the instructions were confusing, which may have led to these inconsistent results.

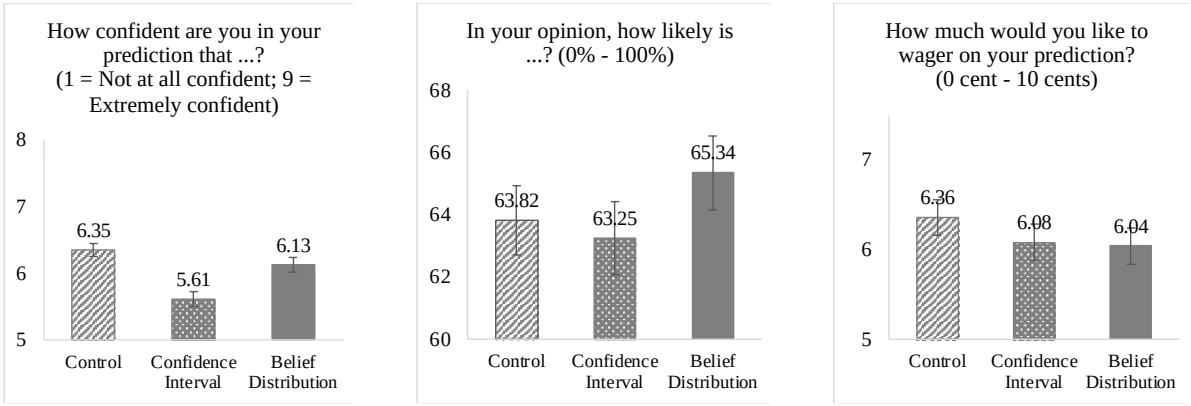

Figure S1. Study S3 Results. Error bars represent +/- 1 clustered standard error.

Supplement 11: Study S4 and S5

To explore why providing a belief distribution might make people more or less confident, we conducted Studies S4 and S5, in which participants provided reasons for increasing or decreasing confidence in a within-subject design. As in other studies in the paper, participants made predictions about other respondents' preferences, or the outcomes of upcoming sports games. We assessed their confidence in the predictions at two time points in the study, the first time after giving their best estimate and the second time after completing the belief distribution task. We asked participants who gave a different response the second time to explain why they felt more or less confident.

Method

Participants. We conducted Studies S4 and S5 using U.S. participants from Prolific. We decided in advance to recruit 200 and 400 participants, respectively. Both studies were not pre-registered. For consistency, we applied the pre-registered exclusion rules for other studies: Only participants who passed the attention check at the beginning of the survey were allowed to proceed. In addition, we excluded participants whose Prolific IDs or IP addresses appeared more than once (9 exclusions and 17 exclusions) and participants who misreported Prolific IDs (2 exclusions and 3 exclusions). In Study S4, this left us with a final sample of 186, averaged 34.8 years of age, and was 47.9% female. In Study S5, this left us with a final sample of 369, averaged 39.4 years of age, and was 26.8% female.

Procedure. We conducted Studies S4 and S5 to explore why eliciting the belief distribution might increase or decrease participants' confidence in their initial predictions. Participants responded to the confidence measures twice in the survey, once immediately after providing the

best estimate, once after providing the belief distribution. Those who provided a different response to either the confidence rating question or the likelihood estimate question provided a free-text explanation of why their confidence rating or likelihood estimate increased or decreased.

The procedure in the two studies was identical with one exception. In Study S4, participants made one prediction about a preference / behavior question, randomly selected from the four questions used in Study 3. In Study S5, participants made two predictions, one for an NFL game, randomly selected from a set of four upcoming games, and one for a preference / behavior question, randomly selected from the same set as in Study S4. The order of the two predictions in Study S5 was counterbalanced.

Results and Discussion

Most participants (over 60% in both studies) reported being equally confident at the two time points. Therefore, when aggregated over all participants, the confidence rating and likelihood estimate before and after giving the belief distribution did not differ significantly (ps > .166 across the two studies).

Among the remaining participants who did change their responses, the number of those who became more confident and the number of those who became less confident were roughly equal (see Table S9). In Study S5 where participants made two predictions, roughly one third of participants adjusted their responses in different directions for the two predictions (125 for confidence rating question, 162 for likelihood estimate question). That is, constructing a belief distribution made the same participant more confident in one prediction but less confident in another prediction. These results suggest that the process of giving a belief distribution could make people simultaneously more and less confident.

Free response coding. We developed the codebook based on an initial review of participants' responses. The full list of codes indicating reasons for increasing (or decreasing) confidence (or likelihood) ratings is listed below.

Reasons for "increasing" confidence or likelihood rating:

1. *The comment implies that giving the entire subjective distribution reinforces the idea that the true answer is close to her estimate (or that other estimates are not very plausible).*
2. *The comment implies that the participant just thought more about it.*
3. *The comment implies that it's gut feeling / intuition.*
4. *The comment suggests another reason.*
5. *The comment gives no reason / is unclear.*

Reasons for "decreasing" confidence or likelihood rating:

1. *The comment implies that giving the entire subjective distribution makes the participant realize there are other possibilities that they didn't think about before.*
2. *The comment implies that the participant just thought more about it.*
3. *The comment implies that it's just gut feeling / intuition.*
4. *The comment suggests another reason.*
5. *The comment gives no reason / is unclear.*

Independent coders blind to the hypothesis applied the coding scheme to all responses from the two studies. The resulting categories are reported in the posted dataset. Table S9 presents a summary of the full qualitative coding results.

Table S9. Free Response Coding in Studies S4-S5

	Study S4, N = 186		**Study S5 (Preference), N = 369**		**Study S5 (NFL), N = 369**	
Categories	Confidence rating	Likehood estimate	Confidence rating	Likehood estimate	Confidence rating	Likehood estimate
Increasing confidence	N = 25	N = 39	N = 45	N = 54	N = 48	N = 52
#1. The comment implies that giving the entire subjective distribution makes the participant realize that the appropriate range contains their best estimate (or that other outcomes are not very plausible).	40.00%	35.90%	24.44%	14.81%	43.75%	21.15%
#2. The comment implies that the participant just thought more about it.	24.00%	25.64%	37.78%	29.63%	20.83%	26.92%
#3. The comment implies that it's gut feeling / intuition.	12.00%	12.82%	26.67%	40.74%	20.83%	26.92%
#4. The comment suggests another reason.	4.00%	0.00%	6.67%	1.85%	6.25%	11.54%
#5. The comment gives no reason / is unclear.	20.00%	25.64%	4.44%	12.96%	8.33%	13.46%
Decreasing confidence	N = 37	N = 31	N = 41	N = 60	N = 35	N = 51
#1. The comment implies that giving the entire subjective distribution makes the participant realize there are other possibilities that they didn't think about before.	29.73%	19.35%	34.15%	21.67%	37.14%	31.37%
#2. The comment implies that the participant just thought more about it.	16.22%	38.71%	21.95%	43.33%	40.00%	35.29%
#3. The comment implies that it's gut feeling / intuition.	35.14%	25.81%	19.51%	23.33%	17.14%	19.61%
#4. The comment suggests another reason.	8.11%	6.45%	14.63%	3.33%	5.71%	11.76%
#5. The comment gives no reason / is unclear.	10.81%	9.68%	9.76%	8.33%	0.00%	1.96%

Among participants who increased confidence, the majority reasoned that giving the belief distribution made them realize their initial prediction was in the appropriate range. For example, a participant stated: *"After I did the part of the survey where it had me rate different percentage brackets on the number of participants who said they liked Thanksgiving more than Christmas, it got me thinking that my guess could very well be in the range of the correct answer."* On the other hand, among those who reduced confidence, a considerable number of responses suggested that the belief distribution reminded them of other previously ignored possibilities. For example, one participant wrote: *"Once I realized how big the scale was, I wasn't as confident anymore. The task with the slider to determine the likeliness of people who prefer the coffee smell made me realize how many different opinions there were and it is hard to generalize."*