Promoting Erroneous Divergent Opinions Increases the Wisdom of Crowds

Autoría ditelliana: Navajas, Joaquín (*Universidad Torcuato Di Tella. Escuela de Negocios. Laboratorio de Neurociencia*)
Otras autorías: Barrera-Lemarchand, Federico; et al.

Fecha de la versión archivada: 11/04/2024

Promoting erroneous divergent opinions increases the wisdom of crowds

Federico Barrera Lermarchand[1,2,3], Pablo Balenzuela[2,3], Bahador Bahrami[4,5,6], Ophelia Deroy[7,8,9], & Joaquin Navajas[1,2,10]

1 Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Buenos Aires, Argentina

2 National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina

3 Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

4 Crowd Cognition Group, Department of General Psychology and Education, Ludwig Maximilian University, Munich, Germany

5 Department of Psychology, Royal Holloway University of London, UK

6 Centre for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

7 Munich Centre for Neuroscience, Ludwig Maximilian University, Munich, Germany

8 Institute of Philosophy, School of Advanced Study, University of London, London, UK

9 Faculty of Philosophy, Ludwig Maximilian University, Munich, Germany

10 Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires, Argentina

Corresponding author: Joaquin Navajas (joaquin.navajas@utdt.edu)

Abstract

The aggregation of many lay judgements generates surprisingly accurate estimates. This phenomenon, called the "wisdom of crowds", has been demonstrated in domains such as medical decision making and financial forecasting. Previous research identified two factors driving this effect: the accuracy of individual assessments and the diversity of opinions. Most available strategies to enhance the wisdom of crowds have focused on improving individual accuracy while neglecting the potential of increasing opinion diversity. Here, we study a complementary approach to reduce collective error by promoting erroneous divergent opinions. This strategy proposes to anchor half of the crowd to a small value and the other half to a large value before eliciting and averaging all estimates. Consistent with our mathematical modeling, four experiments (N=1362 adults) demonstrate that this method is effective for estimation and forecasting tasks. Beyond practical implications, these findings offer new theoretical insights into the epistemic value of collective decision making.

Keywords: wisdom of crowds, collective intelligence, diversity, anchoring, extremization

Research Transparency Statement

Conflicts of interest: All authors declare no conflicts of interest. Funding: This research was supported by the James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition—Scholar Award and by the Templeton World Charity Foundation. Ethics: This research complies with the Declaration of Helsinki (2023) and received approval from a local ethics board (ID: 435-5).

All data and code supporting all analyses in this paper are available at the Open Science Framework (https://osf.io/ch5qw/?view_only=f7f843db75bc4666be5eff771f29cca4).

Experiment 2 was pre-registered at https://aspredicted.org/RYC_4Y5.

Experiment 3 was pre-registered at https://aspredicted.org/HZC_PTH

Experiment 4 was pre-registered at https://aspredicted.org/YQ4_LYS.

Introduction

The aggregation of many lay estimates often outperforms individual expert judgements (De Condorcet, 1785; Galton, 1907). This phenomenon, popularly known as the “wisdom of crowds” (Surowiecki, 2005), has been applied to a wide range of problems such as improving medical diagnoses (Kurvers et al., 2016), forecasting geopolitical events (Mellers et al., 2014), predicting financial markets (Ray, 2010), reverse-engineering the smell of molecules (Keller et al., 2017), and fact-checking news (Allen, 2021), among many others. Given its practical relevance, understanding the conditions under which crowds produce accurate estimates has become a relevant issue in the psychological sciences (Kameda, Toyokawa, & Tindale, 2022; Karachiwalla & Pinkow, 2021; Navajas et al., 2018; Kao & Couzin, 2014).

One important driver of collective accuracy is the diversity of opinions in the crowd (Hong & Page, 2004; Page, 2008; Becker, Porter & Centola, 2019; Shi et al., 2019; Jönsson, Hahn & Olsson, 2015). A simple intuition underlies this claim: when crowds produce diverse estimates, it is likely that some individuals will underestimate the correct answer, while others will overestimate it. Therefore, the more diverse the crowd, the higher the chance that individual errors will cancel out in the aggregation process. More formally, the “Diversity Prediction Theorem” (Page, 2007) states that the crowd’s error (E) can be expressed as the mean individual error (ε) minus the crowd’s predictive diversity (δ, also known as the population variance):

$$E = \varepsilon - \delta \qquad [1]$$

One implication of this mathematical identity (proof can be found in Supplementary Information) is that, in principle, the crowd’s accuracy could be increased either by reducing the individual error (ε) or, alternatively, by increasing the predictive diversity (δ). However, while these two strategies are equally valid in theory, most available studies aiming at increasing the wisdom of crowds have exclusively focused on reducing ε while neglecting the potential of increasing δ.

For example, previous studies have proposed aggregating information from “select” crowds composed by individuals who are more accurate across estimation problems (Mannes, Soll & Larrick, 2014). Other studies have shown that individual error can be reduced by counteracting individual biases (Kao et al., 2018) or by exposing individuals to social information (Jayles et al., 2017; Frey & Van de Rijt, 2021; Madirolas & de Polavieja, 2015; Lorenz et al., 2011). A notable exception to this tendency demonstrated that collective accuracy can be increased by enhancing cognitive-process diversity (Keck and Tang, 2020), a construct which is different from predictive diversity. However, even in that case, it remains unclear whether the method actually increased opinion diversity, reduced mean individual error, or both. A few other works have focused on diversity, but in less direct ways. For example, a previous paper performed secondary analyses to evaluate increases in diversity through the lens of Equation [1], but without directly manipulating diversity or individual accuracy (Nobre & Fontanari, 2020). Similarly, the “crowd within” phenomenon (Vul & Pashler, 2008), where experimenters ask participants to produce more than one estimate, rely on the idea of increasing in diversity within individuals. However, whether and how this procedure increases population diversity remains unknown.

Showing that it is possible to decrease collective error merely by increasing predictive diversity is non-trivial for several reasons. First, it would demonstrate that there are processes that simultaneously increase individual error and collective accuracy. Second, from a practical standpoint, it would provide practitioners with a novel approach to increase collective estimation accuracy. Third, given that the wisdom of crowds has been previously interpreted as empirical evidence for the epistemic value of democratic judgements, this putative dissociation between individual and collective accuracy should mitigate concerns about the increase of misinformed voters in recent elections.

Put together, one converges to a counterintuitive, albeit somewhat uncomfortable possibility: if collective error is reduced by increasing diversity, this would imply that the wisdom of crowds may be enhanced by persuading individuals to adopt erroneous divergent opinions. In this paper, we present theoretical simulations and empirical evidence for this claim. We introduce a new approach to increase collective accuracy by boosting the crowd's predictive diversity at the expense of reducing individual accuracy, even when the truth is completely unknown and unavailable, including to the experimenters.

Increasing the Wisdom of Crowds through Extremization

We propose to promote the adoption of extreme estimates, and therefore to increase diversity, by means of a cognitive bias known as the anchoring effect (Tversky & Kahneman, 1974). The proposed method consists in anchoring one half of the crowd to a small value ("low anchor") and the other half to a large value ("high anchor"), and then averaging all estimates. We hypothesized that this technique should lead to an increase in the predictive diversity that surpasses the increase in mean individual error, thus leading to lower collective error. Using a simple mathematical model, we first demonstrated that this method is expected to enhance collective accuracy across a wide range of parameters. We then empirically tested the procedure across four different experiments and showed that it indeed leads to a substantial reduction of collective error.

Let us consider the scenario where someone needs to estimate a numerical variable that is unknown to them; for example, the height of the Eiffel Tower. Based on the wisdom-of-crowds effect, one could obtain an approximate value by asking a large number of individuals to provide an estimate. Then, to estimate the height of the Eiffel Tower, the person would aggregate these values, for example, by averaging them (**Fig. 1A**). In this work, we propose an alternative approach that consists in dividing the crowd into two halves and extremizing opinions in opposite directions (**Fig. 1B**). We suggest doing so by using the anchoring effect: before estimating the relevant variable, individuals are first asked to consider either an extremely low or high value. In the previous example, half of the individuals would be asked to consider if the height of the Eiffel Tower is greater or less than 10 meters (low anchor, A_L) and the other half, if it is greater or less than 1000 meters (high anchor, A_H). After providing a categorical answer to this initial question, all individuals would then be asked to provide their best-guess estimate. An extensive literature has shown that these estimates should then be consistently biased towards the initially considered values (Furnham & Boo, 2011; Röseler et al., 2022). Because these anchors are extreme in opposite directions, this procedure should extremize the estimates produced by the crowd as a whole, leading to an increase in predictive diversity. We therefore propose to average all numbers, across both halves of the crowd.

While this procedure requires pre-defining two extreme values that will be used as anchors, the strategy does not require knowing the correct answer. However, reasonably, its accuracy will depend on the specific choice of anchors. Therefore, to understand better the conditions under which the proposed approach is expected to increase collective accuracy, we developed a simple mathematical model of the anchoring effect.

Model

We consider a set of individuals who, being asked to estimate the variable θ, produce a distribution of values with mean μ. The model assumes that, when those individuals are anchored to a low value A_L, they produce a set of estimates with a different mean

$$\mu_L = w_L\, A_L + (1 - w_L\,)\, \mu, \qquad [2]$$

where $0 \leq w_L \leq 1$ is an anchoring weight reflecting the strength of the anchoring procedure given by the anchor A_L. Similarly, a population anchored to a high value A_H would produce a distribution of estimates given by

$$\mu_H = w_H\, A_H + (1 - w_H\,)\, \mu, \qquad [3]$$

where $0 \leq w_H \leq 1$ is the corresponding anchoring weight given by anchor A_H.

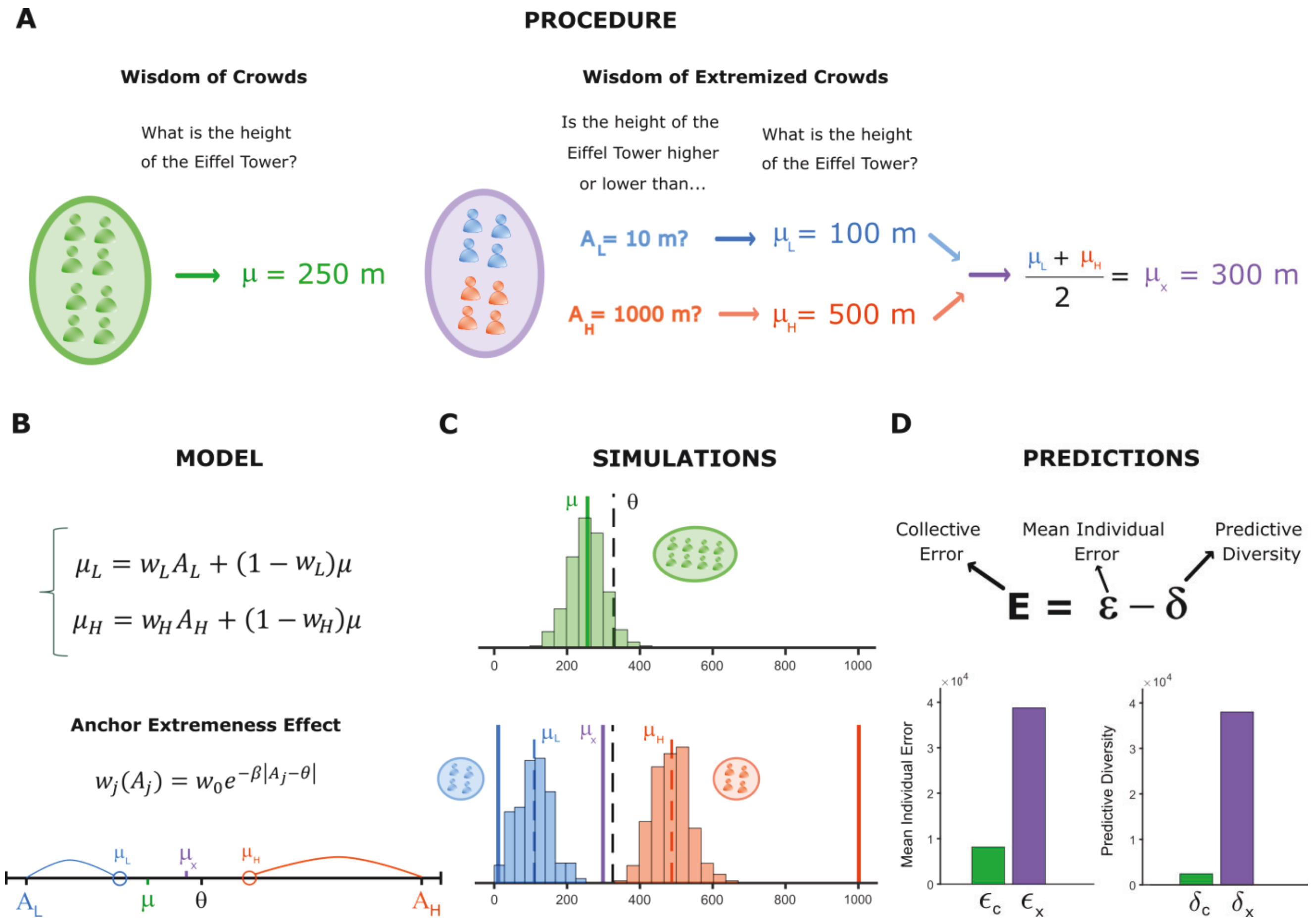
A
PROCEDURE
Wisdom of Crowds
What is the height of the Eiffel Tower?
μ = 250 m
Wisdom of Extremized Crowds
Is the height of the Eiffel Tower higher or lower than…
What is the height of the Eiffel Tower?
A_L = 10 m?
μ_L = 100 m
A_H = 1000 m?
μ_H = 500 m
(μ_L + μ_H)/2 = μ_x = 300 m
B
MODEL
μ_L = w_L A_L + (1 − w_L)μ
μ_H = w_H A_H + (1 − w_H)μ
Anchor Extremeness Effect
w_j(A_j) = w_0 e^{−β|A_j−θ|}
A_L
μ_L
μ
μ_x
θ
μ_H
A_H
C
SIMULATIONS
μ
θ
0
200
400
600
800
1000
μ_L
μ_x
μ_H
D
PREDICTIONS
Collective Error
Mean Individual Error
Predictive Diversity
E = ε − δ
×10^4
Mean Individual Error
0
1
2
3
4
ϵ_c
ϵ_x
Predictive Diversity
δ_c
δ_x

Fig. 1. Wisdom of extremized crowds. (A) The method for estimating a variable through the wisdom of crowds consists in asking a crowd of individuals to estimate a given quantity, and averaging all answers (averages represented by μ). The proposed alternative method consists in promoting the adoption of divergent opinions within a crowd by dividing it in two halves, asking an anchoring question with either a low (L) or a high (H) value to each half, and then averaging all answers (μ_x). **(B)** Mathematical model of the anchoring effect. The anchored mean is a weighted average of the anchor and the wisdom-of-crowds value. The weight w depends on the difference between the anchor and the correct answer, reflecting an internal sensitivity to the correct answer. **(C)** Simulations performed using the proposed model show that the method is expected to outperform the wisdom of crowds. **(D)** Mean individual error and predictive diversity on simulated data show an increase in both individual error and predictive diversity. However, the increase in individual error (left panel) is smaller than the increase in predictive diversity (right panel), resulting in an overall reduction in collective error ("c" represents the control group, and "x" the extremized crowd).

Eq. [2] and [3] imply that the anchored mean (either low or high) is a weighted average of the anchor and the mean of the original distribution of values μ. Following a variety of empirical findings linking the anchoring effect to the plausibility of the anchor (Mussweiler & Strack, 2001; Wegener et al., 2001), we propose that the weights w are given by

$$w_j = w_0 e^{-\beta|A_j-\theta|} \quad [4]$$

where j indicates whether the weight corresponds to the low or high anchor (w_L or w_H, respectively) and w_0 is a parameter reflecting the anchoring weight when individuals are anchored to the correct value θ. The parameter β is an "inverse temperature" encoding the sensitivity of the individuals to the distance between the anchor and θ. Thus, the value of β modulates the strength of the "anchoring extremeness effect" (Röseler et al., 2022).

In this work, we propose averaging estimates from two populations of individuals, each of which is anchored to either a low or a high value (A_L or A_H). Assuming both populations are equally sized, we can estimate the mean estimates from both populations as

$$\mu_x = \frac{\mu_L+\mu_H}{2} \quad [5]$$

Simulations (for a set of parameters, with θ=324, μ=250, A_L=10, A_H=1000, w_0=1, and β=0.0017) show that this model is expected to produce a reduction in collective error (**Fig. 1C**). This increase in collective accuracy is due to an increase in diversity that surpassed the observed reduction in individual accuracy (**Fig. 1D**). The combination of increased diversity accompanied by a reduction in individual accuracy is a general tradeoff appearing under different model specifications (**Fig. S1**).

The method implicitly assumes that it is possible to select anchors in a way in which the correct value will be underestimated by the low anchor and overestimated by the high anchor. However, meeting this condition is neither sufficient nor necessary. For example, the method should also work when both anchors have the opposite bias than the non-anchored responses. Therefore, to understand the conditions under which the method is expected to work, we performed a fine-grained model-based analysis.

Analytically, we found that the key variable determining the success of the approach is the mid-point of the anchors, defined as $\bar{A} = \frac{A_L + A_H}{2}$. Following a simple mathematical procedure (for details, see Supplementary Information), we found that the range (Δ) of $\bar{A}$ values where the method outperforms the wisdom of crowds is

$$\Delta = \frac{4|\mu - \theta|}{w_L + w_H} \qquad [6]$$

The expression derived in Eq. [6] implies that the range of values where the method outperforms the wisdom of crowds is always equal to or larger than two times the collective error. This can be shown by examining two opposite extreme scenarios. If the sensitivity β is small (i.e., when the anchoring procedure does not depend on the distance between the anchor and the truth), the range of values where the method outperforms the wisdom of crowds converges to, at the very least, twice the collective error (i.e., if $\beta \rightarrow 0$, then $w_j \rightarrow w_0$, and therefore $\Delta \rightarrow \frac{2|\mu - \theta|}{w_0}$). In the opposite case, when the sensitivity β is large (i.e., when the anchoring effect is stronger as anchors become closer to the correct answer), then this method is always better than the wisdom of crowds (i.e., if $\beta \rightarrow \infty$, then $w_j \rightarrow 0$, and thus $\Delta \rightarrow \infty$). To evaluate empirically the effectiveness of the proposed method, we performed four experiments.

Experiment 1: Estimation of unbounded quantities

Methods

Participants and Questions

In Experiment 1, N=120 participants (48 female, aged 37.2 ±11.6 yr, from the USA, recruited online on Amazon Mechanical Turk) provided estimates about 14 general-knowledge quantities (**Table S3**). All variables were positive and unbounded, like the example used in **Figure 1** (e.g., "how many bridges are there in Paris?"). Participants had monetary incentives to estimate these variables as accurately as possible. Participants were informed that their participation was completely voluntary, and that they could withdraw their participation at any time. All data were completely anonymous. The experimental protocol was approved by a local ethics committee (anonymized for peer review purposes).

Procedure

This experiment was developed using Psytoolkit (Stoet 2010; Stoet 2017). One third of the sample was randomly assigned to a control condition where they simply estimated a variable. The other two thirds of the participants were randomly assigned to the experimental condition where, before estimating the quantity, they were asked to consider either an extremely low or extremely high value (e.g., is the number of bridges in Paris higher or lower than 349?). Half of the anchored participants considered a low value, and the other half considered a high value (randomly assigned). Crucially, these extreme values were not manually chosen by the experimenters, but set automatically as the 5 and 95 percentiles of the empirical values observed in the control condition (to this aim, the data from the first third of the sample was collected prior to the remaining data). Overall, we collected data from 41 participants in the control condition and 79 participants in the experimental condition. Our motivation for collecting this number of participants per condition was related to the maximum crowd size which could be obtained for our resampling procedure (see "Data Analysis" below). We aimed for a crowd size of at least 30. In all cases, the questions were randomly ordered. All participants had a maximum of 15 seconds to answer. Participants were paid a flat fee of 1.5 USD for their participation. Estimation accuracy was incentivized by rewarding a bonus payment of 0.5 USD to the top 10% most accurate respondents.

Data Analysis

We discarded data from participants that completed the survey in less than three minutes, or those who failed to complete the survey. We also excluded participants with two or more exactly correct answers (which is likely to reflect cheating).

To compare different conditions in a properly balanced way, we developed a resampling bootstrapping strategy. For each crowd size, we randomly selected with replacement a fixed number of individuals and estimated the collective error, predictive diversity, and mean individual error for that crowd size and that iteration. We repeated that procedure 1,000 times for crowd sizes that varied from 2 to different maximum values in different experiments (i.e., n=32 in Experiment 1, n=50 in Experiment 2, n=100 in Experiment 3, and n=70 in Experiment 4). In all cases those values allow perceiving the asymptotic behavior of the collective error as a function of crowd size.

Results and Discussion

By employing a simple bootstrapping resampling method, we estimated the collective error of differently-sized groups for both the wisdom of crowds and the wisdom of extremized crowds (**Fig. 2A**). We observed that the average collective error of the latter was always smaller than the former. For example, the collective error of 34 individuals randomly taken from the control condition was substantially larger than the collective error of 34 extremized individuals (unpaired t-test: t(998)=29.7, p=2x10^{-139}; effect size: Cohen's d = 1.8 ± 0.1, 95% CL). This collective error reduction was due to an increase in predictive diversity (**Fig. 2B**, unpaired t-test: t(998)=54.8, p<10^{-200}; effect size: Cohen's d = 3.3 ± 0.1, 95% CL) that was higher than the increase in mean individual error (**Fig. 2B**, unpaired t-test: t(998)=24.2, p=5x10^{-102}; effect size: Cohen's d = 1.4 ± 0.1, 95% CL).

One limitation of Experiment 1 is that the anchors were defined after collecting the data of the non-anchored population. Because this procedure may be inconvenient from a practical point of view, we performed a second pre-registered experiment with bounded quantities, where anchors were pre-defined and fixed across all questions (https://aspredicted.org/RYC_4Y5).

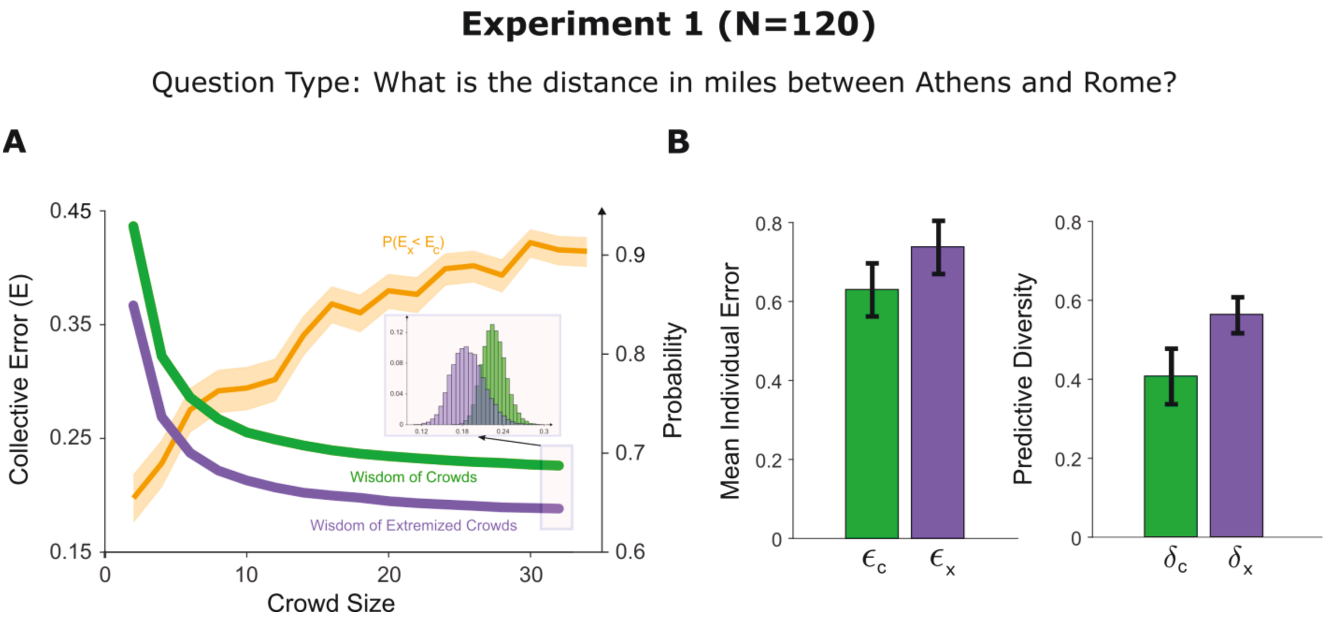

Fig. 2. Empirical results for Experiment 1. (A) Collective error as a function of crowd size, for the non-extremized wisdom of crowds (green) and for the wisdom of extremized crowds (purple). The standard error of each curve is within the line width. The inset shows the distribution of values from the resampling method for the largest crowd size (N=34). The orange line depicts the probability of a sample from the distribution of collective errors of the extremized crowd (purple distribution, E_x) being smaller than a sample from the distribution of collective errors of the non-extremized crowd (green distribution, E_c). **(B)** Mean individual error and predictive diversity for both the non-extremized wisdom of crowds (green) and the wisdom of extremized crowds (purple). The error bars are the standard deviation of the means.

Experiment 2: Estimation of bounded quantities with fixed anchors

Methods

Participants and Questions

In Experiment 2, we recruited N=396 participants online (235 female, aged 27.9±8.8 yr), and asked them 30 general-knowledge questions that involved the estimation of a percentage. Therefore, all answers were bounded in the range [0,100] (e.g. what percentage of the population of Argentina is under 15 years old? See Table S4 for the full list of questions used in Experiment 2). Unlike the previous study, here we used the same anchors for all questions, always set at either 5% (low anchor A_L) or 95% (high anchor A_H). Participants were recruited online, and resided in Argentina at the time of the experiment. All questions came from a variety of representative surveys carried out by official entities. The contents of the questions were explicitly related to Argentina, and ranged from demographic (e.g. what percentage of the population over 20 years old is either overweight or obese?) to personal perceptions of the country (e.g., what percentage of the population believes abortion is not morally acceptable?). When necessary, we referenced the year in which the corresponding survey had taken place.

Procedure

This experiment was also developed using Psytoolkit (Stoet 2010; Stoet 2017). In order to reduce the length of the survey, we divided the questions into two subsets of 15 questions. Each participant was randomly assigned to one of these sets of questions. In addition, following a very similar procedure to the previous experiment, we randomly divided the sample in two groups. Participants in the control condition were asked to directly estimate the answer to the 15 questions (e.g., What percentage of the population of the USA is under 18 years old?). Participants in the experimental condition answered the same questions but after an "anchoring" question (Do you think the percentage of the population of the USA under 25 years old is above or below 95 %?). The assignment to the low anchor or high anchor condition was random across questions. Overall, we collected data from 119 participants in Group 1, and 277 participants in Group 2. In all cases, the questions were randomly ordered (within each subset of 15 questions). In this case, we estimated the number of participants per condition required from power analyses derived from our previous experiment (as per our pre-registration, https://aspredicted.org/RYC_4Y5). All participants had 20 seconds to answer the questions. Estimation was not incentivized for accuracy.

Data Analysis

For this experiment we followed the same exclusion criteria as for the previous one. In this case, the exclusion criteria were pre-registered. For Experiment 2 we followed the same bootstrapping procedure as stated for our previous experiment.

Results and Discussion

We observed very similar results to Experiment 1 (**Fig. 3A and 3B**). Collective error was lower for the extremized crowd (crowd size N=50, unpaired t-test: t(998)=19.1, p=2x10^{-69}; effect size: Cohen's d = 1.2 ± 0.1, 95% CL). This was accompanied by an increase in mean individual error (unpaired t-test: t(198)=121.1, p<10^{-200}; effect size: Cohen's d = 7.6 ± 0.3, 95% CL) as well as an increase in predictive diversity (unpaired t-test: t(198)=162.5, p<10^{-200}; effect size: Cohen's d = 10.3 ± 0.3, 95% CL).

Given that this experiment used fixed anchors across all questions, it allowed us to test one key element of the model, i.e., the anchoring extremeness effect (Eq. [4]). We did so by performing two separate analyses. First, we examined the biases associated with each experimental condition. We reasoned that, if anchoring was sensitive to the distance to the correct answer, then the effectiveness of the procedure should be higher when the correct answer is close to the anchor. For example, we should see that participants considering a low value (5%) should be more attracted to the anchor when the correct answer is low (below

50%) compared to when the correct answer is high (above 50%). Consistent with this idea, when the correct answer was below 50% (17 questions), we did not observe any statistical difference between the distribution of estimates provided by the population considering the "low anchor" and the correct answer, (paired t-test, t(16)=0.95, p=.36, effect size: Cohen's d = 0.23 ± 0.67, 95% CL), and Bayes factor analysis provided moderate support for the null hypothesis (Bayes factor = 2.7). More importantly, both the non-anchored (paired t-test, t(16)=4.16, p=3×10^{-4}, effect size: Cohen's d = 1.11 ± 0.72, 95% CL) and the population anchored to a high value (paired t-test, t(16)=5.99, p=2×10^{-5}, effect size: Cohen's d = 1.45 ± 0.76, 95% CL) provided a distribution of estimates that significantly overestimated the correct answer (**Fig. 3C**). Conversely, when the correct answer was above 50% (13 questions), we observed the opposite pattern: the population considering the "high anchor" provided a distribution of estimates that was not statistically distinct from the correct answer (paired t-test, t(12)=2.01, p=0.07, effect size: Cohen's d = 0.56 ± 0.78, 95% CL), and Bayes factor analysis suggested very weak evidence in favor of the alternative hypothesis (Bayes factor = 0.77). Crucially, both the non-anchored population (paired t-test, t(16)=3.90, p=0.001, effect size: Cohen's d = 1.08 ± 0.82, 95% CL) and the population anchored to a low value (paired t-test, t(16)=4.90, p=4×10^{-4}, effect size: Cohen's d = 1.36 ± 0.85, 95% CL) provided a distribution of estimates that significantly underestimated the correct answer.

Second, we directly examined the anchoring extremeness effect by studying the association between the strength of the anchoring procedure and the distance between the anchor and the correct answer (Eq. 4). We estimated the strength of the anchoring effect by calculating the anchoring weight, which has very similar properties to the "anchoring index" from previous literature (Jacowitz & Kahneman, 1995). Both quantities (see Eq. 7 in Supplementary Information for details) take a value of 0 when the anchoring procedure does not produce any effect on the estimates, and a value of 1 when the estimates are on average equal to the anchor. Consistent with the existence of the anchoring extremeness effect (Roseler et al., 2022), we observed a significantly negative correlation between the anchoring weight and the distance between the anchor and the correct answer (Pearson correlation coefficient, r=-0.56, p=3x10^{-6}). This empirical observation provides support to the proposed model of the anchoring effect used in Eqs 2-4.

To test the functional form of the proposed model, we compared it with two other monotonically decreasing functions. One is a linear function and the other one is a general hyperbolic function

$$w = \frac{w_0}{1+\beta|A-\theta|}, \quad [7]$$

which is similar to the one used in the temporal discounting literature (Ruggeri et al., 2022). We found that the exponential model led to better fits to the data as assessed by both the Akaike and Bayesian Information Criteria (Linear model: ΔAIC = 1.9, ΔBIC = 1.9; Hyperbolic model: ΔAIC = 0.59, ΔBIC = 0.59).

Another assumption of the proposed anchoring model is that low and high anchors elicit the same anchoring weights. Given that this feature of the model could be an oversimplification for specific problems (e.g., city populations), we compared the goodness of the fits achieved with a single anchoring weight per question, with the one obtained with different weights for low and high anchors. Formal model comparison indicated that having a single weight per question led to better fits than the model with asymmetrical weights (ΔAIC = 1.6, ΔBIC = 5.7).

Finally, this experiment allowed us testing if the proposed method is robust to the use of different aggregation procedures like the simple average, the median, and a performance-weighted average (Mannes et al., 2014; Collins et al, 2023). We found evidence that, regardless of the specific aggregation procedure, crowds extremized through anchoring were always more accurate than non-extremized crowds (Fig. S3; for details, see Supplementary Information).

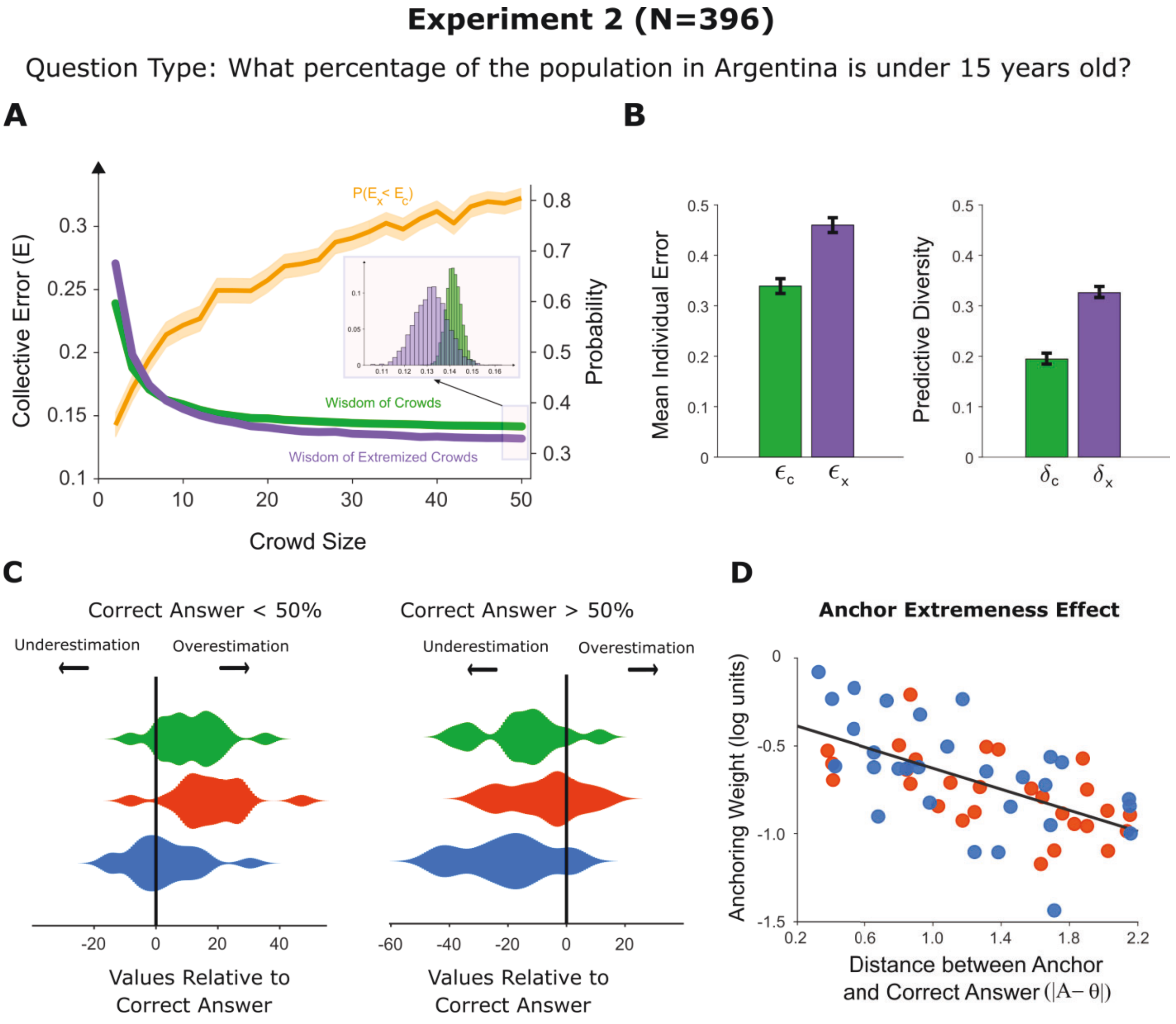

Fig. 3. Empirical results for Experiment 2. (A) Collective error as a function of crowd size, for the non-extremized wisdom of crowds (green) and for the wisdom of extremized crowds (purple). The standard error of the curves is within the line width. The inset shows the distribution of values from the resampling method for the largest crowd size (N=50). The orange line depicts the probability of a sample from the distribution of collective errors of the extremized crowd (purple distribution, E_x) being smaller than a sample from the distribution of collective errors of the non-extremized crowd (green distribution, E_c). **(B)** Mean individual error and predictive diversity for both the non-extremized wisdom of crowds (green) and for extremized crowds (purple). The error bars show the standard deviation of the means. **(C)** Distributions of values corresponding to the difference between the mean answers and the correct answer for each question, for the non-extremized wisdom of crowds (green), the crowd extremized using

a high anchor (red), and the crowd extremized using a low anchor (blue). We separate the cases where the correct answer is above 50% (left panel) and where it is above 50% (right panel). The black line depicts the case where the mean value is equal to the correct answer. **(D)** Empirical anchoring weight (*w*) for each question. Blue dots show estimates using low anchors and red dots show the same with high anchors. The horizontal axis represents the distance between the corresponding anchor and the correct answer (logarithmic units on anchoring weight), and the black line shows the best linear fit of the data.

Experiment 3: Forecasting

Subsequently, we asked whether the proposed method can be useful to outperform the wisdom of crowds on forecasting tasks, i.e., for domains where the truth is unknown and unavailable at the time of the experiment. To answer this question, we performed a third pre-registered experiment that took place in the midst of the COVID-19 crisis (https://aspredicted.org/HZC_PTH).

Methods

Participants and Questions

We recruited N=620 participants (312 female, aged 46.1 ± 15.7 yr.) from the USA, and asked them to estimate the total number of COVID-19 cases and deaths that would occur in the United States in the following week (from 27 July to 2 August, 2020). Thus, the answers were positive (unbounded), resembling a typical wisdom-of-crowds experiment, but related to quantities that were unknown at the time of the experiment. Participants were recruited online by using Prolific (https://www.prolific.co/), and resided in the United States at the time of the experiment. Participants had monetary incentives to estimate these variables as accurately as possible. Anchors were selected as extreme values based on historical data, namely, two orders of magnitude less or more the number of COVID-19 cases and deaths reported in the two weeks before the beginning of the experiment (see Table S4 for the questions used in Experiment 3).

Procedure

In Experiment 3, developed using Survey Monkey (https://www.surveymonkey.com/), we asked two questions related to the COVID-19 pandemic: we asked participants to forecast the number of COVID-19 deaths and cases in the week following the experiment. Anchors were selected as extreme values based on historical data, namely, two orders of magnitude less or more the number of COVID-19 cases and deaths reported in the two weeks before the beginning of the experiment. The correct answers to those questions were unknown at the time of the experiment. Since all participants answered both questions, there were a total of six conditions in this experiment. In Condition 1, we used a "low anchor" on COVID-19 deaths. We asked participants whether they thought there would be more or less than 40 new deaths in the week following the experiment, and then asked them to forecast the number of new deaths. In the next screen, participants also forecast the number of new COVID-19 cases in the week following the experiment). Condition 2 was the same as Condition 1, but with a "high anchor" (400,000) on COVID-19 deaths. In Condition 3, we did not anchor people's expectations (which served as a control for Conditions 1 and 2). We first asked people to forecast the number of new deaths in the following week. In the following screen, participants forecast the number of new COVID-19 cases in the same week. Condition 4 was analogous to Condition 1, but changing the order of questions and setting an anchor on cases, instead of deaths. First, we asked participants whether they thought there would be more or less than 8,000 new cases in the following week, and then asked them to estimate the number of new cases. In the next screen, participants forecast the number of new COVID-19 deaths in the same week. Condition 5 was the same as Condition 4, but with a "high anchor" (8,000,000) on COVID-19 cases. In Condition 6, we did not ask any anchoring questions and participants were directly asked to forecast the number of new cases and then the number of new deaths in the week following the experiment. This condition was analogous to Condition 3, but changing the order of questions, and serves as a control for Conditions 4 and 5.

We collected data from 117 participants in Condition 1, 105 participants in Condition 2, 97 participants in Condition 3, 92 participants in Condition 4, 97 participants in Condition 5, and 112 participants in Condition 6. The total sample of 600 participants, obtained through Prolific (https://www.prolific.co/), was representative of the US population in terms of age, gender and ethnicity. In this case, we aimed for 100 participants per condition, in order to have a maximum crowd size of 100 for our resampling analysis (as was pre-registered at https://aspredicted.org/HZC_PTH). This was done to further assess the asymptotic behavior of the collective errors with crowd size. All participants received a flat participation fee of 1.0 USD and, to incentivize forecasting accuracy, we paid a bonus of 2.0 USD to the top 10% performers. There was no time limit to answer these questions.

Data Analysis

For Experiment 3, given that it consisted in forecasting questions, and since every participant completed the survey, there was no need to exclude any of them. All of them answered both forecasting questions. We followed the same bootstrapping procedure as stated for our previous experiments.

Results and Discussion

Again, as in Experiments 1 and 2, we observed that the collective error was lower for the extremized crowd for all group sizes compared to the non-extremized wisdom of crowds (**Fig. 4A**). We found that the decrease in collective error (specifically, for the largest crowd size N=100, unpaired t-test: t(998)=16.4, p=7x10^{-54}; effect size: Cohen's d = 1.04 ± 0.09, 95% CL) was due to an increase in predictive diversity (crowd size N=100, unpaired t-test: t(998)=37.6, p=7x10^{-193}; effect size: Cohen's d = 2.5 ± 0.1, 95% CL). This increase in predictive diversity was in turn greater than the increase in mean individual error (**Fig. 4B,** crowd size N=100, unpaired t-test: t(998)=18.1, p=1x10^{-64}; effect size: Cohen's d = 1.2 ± 0.1, 95% CL). Overall, this study demonstrates that the presented strategy is also useful for a forecasting task.

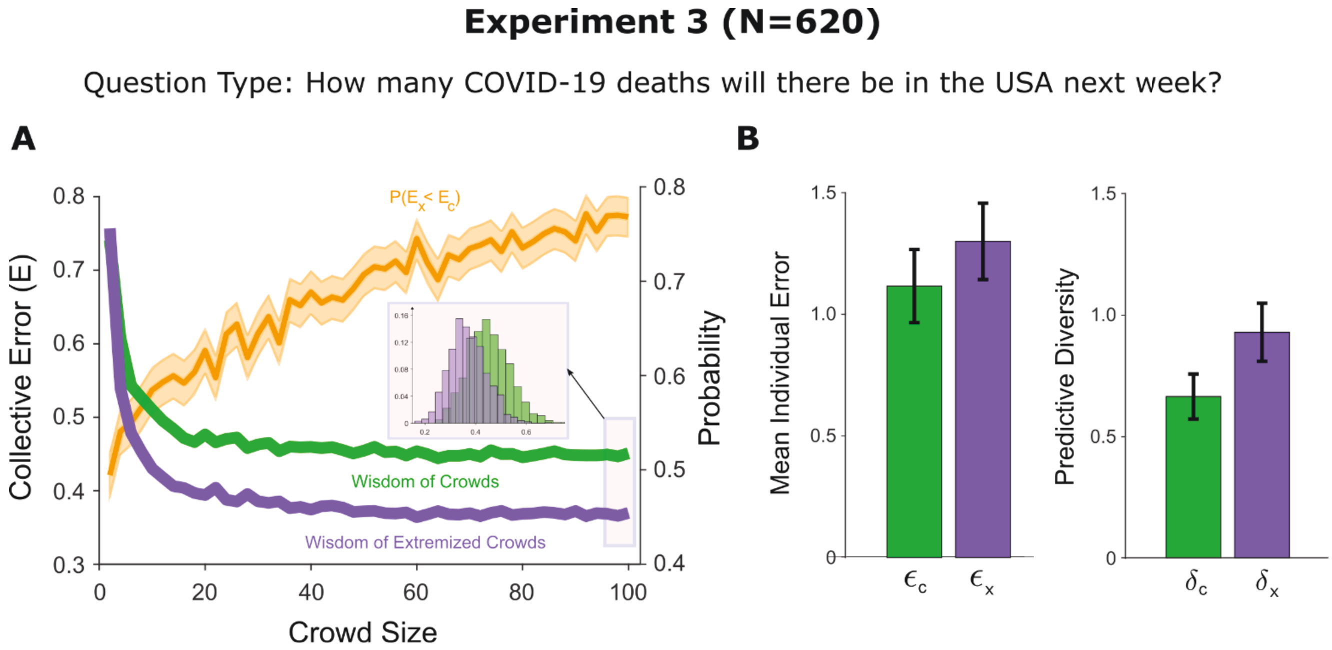

Fig. 4. Empirical results for Experiment 3.

(A) Collective error as a function of crowd size, for the non-extremized wisdom of crowds (green) and for the wisdom of extremized crowds (purple). The standard error of the curves is within the line width. The inset shows the distribution of values from the resampling method for the largest crowd size (N=100). The orange line depicts the probability of a sample from the distribution of collective errors of the extremized crowd (purple distribution, E_x) being smaller than a sample from the distribution of collective errors of the non-extremized crowd (green distribution, E_c). **(B)**. Mean individual error and predictive diversity for both the non-extremized wisdom of crowds (green) and the extremized crowd (purple). The error bars are the standard deviation of the means.

Experiment 4: Self-generated anchors

One limitation of Experiments 1-3 is that, in all cases, participants had access to additional information (either because anchors were pre-defined by the experimenters or based on the responses given by other participants). Experiment 4 addressed that limitation by using self-generated anchors. The experiment was pre-registered (https://aspredicted.org/YQ4_LYS), and used questions similar to those tested in Experiment 1. The key difference is that participants were prompted to generate their own anchors.

Methods

Participants and Questions

We recruited N= 226 participants (110 female, aged 37.2± 12.1 yr, residents of Argentina at the time of the experiment, recruited online) and asked them 10 general-knowledge questions that involved the estimation of the height of a building (e.g. what is the height of the Eiffel Tower?). Unlike all the previous experiments, the anchors were self-generated, meaning that participants provided their own respective anchors (See Table S5 for the full list of questions).

Procedure

This experiment was developed using Psytoolkit (Stoet 2010; Stoet 2017). One third of the sample was randomly assigned to a control condition, where they simply estimated a variable. After participants completed all ten questions, we asked them to provide new, different estimates for each one of them. This procedure allowed us comparing whether the self-generation of anchors in the treatment conditions was different or not to the phenomenon of dialectical bootstrapping (Herzog & Hertwig, 2009)

The two other thirds of the participants were randomly assigned to the experimental condition where, before starting the main questionnaire, they were asked to estimate the height of the tallest building on Earth and the height of the shortest building on Earth. The values they provided were used as the high and low anchors, respectively, for all the estimation questions. We added 10% zero-mean uniform noise to these preliminary estimates, so that each question with the same base anchor had a slightly different value. Thus, before estimating the quantity in question, they were asked to consider either an extremely low or an extremely high value (e.g., is the height of the Eiffel Tower higher or lower than 1000 meters?). For each question, half of the anchored participants considered a low value, and the other half considered a high value (randomly assigned). Overall, we collected data from 75 participants in the control condition and 151 participants in the experimental condition. The sample size for each condition was determined through power analyses (as per our pre-registration, for details see https://aspredicted.org/YQ4_LYS). In all cases, the questions were presented in random order. All participants had a maximum of 25 seconds to answer. Estimation was not incentivized for accuracy.

Data Analysis

We discarded data from participants that completed the survey in less than two minutes, or those who failed to complete half or more of the questions in the survey. We also excluded participants with two or more exactly correct answers (which likely reflects cheating). These were the same exclusion criteria as for Experiments 1 and 2. These exclusion criteria were pre-registered. For Experiment 4 we followed the same bootstrapping procedure as stated for our previous experiments.

Results and Discussion

As with previous experiments, we observed that collective error was lower for the extremized crowd for all group sizes compared to the non-extremized wisdom of crowds (**Fig. 5A**). We found that the decrease in collective error (specifically, for the largest crowd size N=70, unpaired t-test: t(998)=58.4, $p<10^{-200}$; effect size: Cohen's d = 3.7 ± 0.1, 95% CL) was due to an increase in predictive diversity (crowd size N=70, unpaired t-test: t(998)=78.5, $p<10^{-200}$; effect size: Cohen's d = 4.9 ± 0.2, 95% CL). This increase in predictive diversity was in turn greater than the increase in mean individual error (**Fig. 5B**, crowd size N=70, unpaired t-test: t(998)= 33.9, p=2x10^{-168}; effect size: Cohen's d = 2.2 ± 0.1, 95% CL). Overall, this study demonstratesd that the presented strategy is also useful when employing self-generated anchors. This means that the problem of choosing appropriate anchors can be easily solved by asking participants to provide their own anchors.

We then studied the relationship between this procedure and the wisdom-of-the-inner-crowd effect. Given that participants generated their own anchor, we reasoned that this effect could , in principle, partially explain the efficacy of the method. However, we were able to rule out this possibility, as the wisdom of extremized crowds produced much more accurate collective estimates than the wisdom of the inner crowd (for details, see "Self-generated Anchors and the Wisdom of the Inner Crowd" in Supplementary Information).

Experiment 4 (N=226)

Question Type: What is the Height of the Eiffel Tower? (with self-generated anchors)

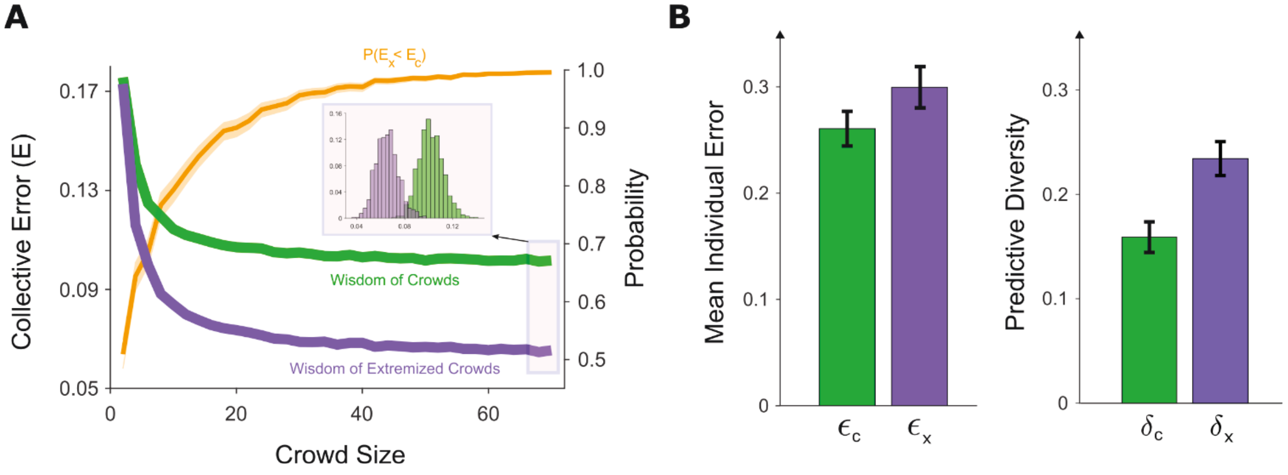

Fig. 5. Empirical results for Experiment 4.

(A) Collective error as a function of crowd size, for the non-extremized wisdom of crowds (green) and the wisdom of extremized crowds (purple). The standard error of the curves is within the line width. The inset shows the distribution of values from the resampling method for the largest crowd size (N=70). The orange line depicts the probability of a sample from the distribution of collective errors of the extremized crowd (purple distribution, E_x) being smaller than a sample from the distribution of collective errors of the non-extremized crowd (green distribution, E_c). **(B)** Mean individual error and predictive diversity for both the non-extremized wisdom of crowds (green) and the extremized crowd (purple). The error bars are the standard deviation of the means.

General Discussion

In this work, we introduce a novel strategy to outperform the wisdom of crowds (Surowiecki, 2005) that has both practical and theoretical implications. By means of the anchoring effect (Tversky, & Kahneman, 1974), we show we can reduce collective error by increasing predictive diversity (Page, 2007). In previous literature, methods to increase the wisdom of crowds often involved strategies to reduce individual error (Madirolas & de Polavieja, 2015; Mannes, Soll & Larrick, 2014). However, while theoretical analysis suggested this goal could also be achieved by increasing the predictive diversity (as seen in Eq. [1]), that possibility remained unexplored. Here, we thoroughly studied this approach both theoretically and empirically: first, by developing a mathematical model of the anchoring effect and, second, by performing four behavioral experiments.

In all of the experiments, regardless of differences in sample size, country of implementation, use of bounded or unbounded quantities, anchor-setting procedure, and whether the task involved estimation or forecasting, we observed very similar results. Extremized crowds always produced estimates with lower collective error, and this was always accompanied by an increase in both mean individual error and predictive diversity. This demonstrates that it is possible to increase collective intelligence while concurrently reducing individual accuracy, an approach that remained heretofore empirically untested. Therefore, this should inspire future research aimed at increasing the wisdom of crowds using a similar strategy.

One limitation of the proposed method is that the selection of appropriate anchors could potentially prove hard (for example, in forecasting problems). However, our model-based analyses as well as our empirical findings suggest that the range of values where this method improves the wisdom of crowds is large (Equation 4). We also showed that adequately selecting anchors is empirically feasible across four very different setups. In Experiment 1, the anchors were chosen based on extremely high and low values (5 and

95 percentiles) of the non-anchored distribution. This procedure, however, has the disadvantage that it shares information between different groups of participants and this could partly explain changes in collective performance (e.g., Navajas et al., 2018; Becker et al., 2017; Becker et al., 2021). In Experiment 2, we studied the estimation of percentages while keeping anchors fixed at 5% and 95%. With this experiment we showed that the method still works in a condition where anchors were pre-defined across all questions. This setup also allowed us obtaining empirical support for the proposed anchoring model. In Experiment 3, we shifted the focus of attention to forecasting problems. Anchors were defined a priori, using historical data. Given that the correct answer was unknown to the experimenters before conducting the study, this setup demonstrated that -even in highly uncertain forecasting problems- it is feasible to select anchors so that the method would still work. However, one limitation of Experiments 1-3 is that, in all cases, participants had access to additional information (either because anchors were pre-defined by the experimenters or based on the responses given by other participants). Experiment 4 addressed that limitation by using self-generated anchors. This demonstrates that this method is useful even in conditions where anchors are not selected by the experimenters.

The first study on the wisdom of crowds (Galton, 1907) was regarded as an empirical demonstration that democratic aggregation principles can be reliable and efficient. This was counterintuitive at the time, since it showed that erroneous individuals could make good collective choices. Nowadays, when political opinions tend to become extremized, these results seem to suggest that democratic decisions can still be surprisingly accurate, even if collective choices proceed from misinformed voters, as long as they are sufficiently diverse. Therefore, one interpretation of these findings is that opinion polarization, which may stem from the attraction towards political extremes (Goldenberg et al., 2023; Zimmerman et al., 2022), can potentially improve democratic judgement, as previously proposed (Shi et al., 2019).

Studies examining how diversity influences group performance have provided dissimilar results, with both positive (e.g., Mohammed & Ringseis, 2001) and negative (e.g., Navajas et al., 2022) effects present in the collective decision-making literature (for a review, see Sulik, Bahrami & Deroy, 2022). This suggests that accumulating diverse evidence (Couch, 2022) may not be good or bad *per se*, but depend on the interaction with variables such as group size (Pescetelli, Rutherford & Rahwan, 2021) and network structure (Baumann, Czaplicka, & Rahwan, 2024). This paper adds to this literature by showing that increasing diversity may reduce accuracy at the individual level but increase it at the collective level. Crucially, we demonstrated this novel insight in a complex forecasting domain, where there is a longstanding tradition examining the epistemic value of combining diverse predictions (Bates & Granger, 1969). Although this work demonstrated the existence of this phenomenon and studied its robustness across different aggregation methods, we believe future work can refine the wisdom of extremized crowds, for example, by combining them with the use of debiasing through generative models (Lee, & Danileiko, 2014; Lee et al., 2023).

References

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science advances*, *7*(36), eabf4393.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the operational research society*, 20(4), 451-468.

Baumann, F., Czaplicka, A., & Rahwan, I. (2024). Network structure shapes the impact of diversity in collective learning. *Scientific Reports*, 14(1), 2491.

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26), E5070-E5076.

Becker, J. A., Guilbeault, D., & Smith, E. B. (2022). The crowd classification problem: Social dynamics of binary-choice accuracy. Management Science, 68(5), 3949-3965.

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22), 10717-10722.

Couch, N. (2022). The diversity principle and the evaluation of evidence. *Psychonomic Bulletin & Review*, 29(4), 1270-1294.

De Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'impremerie royale.

Frey, V., & Van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management science*, 67(7), 4273-4286.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, *40*(1), 35-42.

Galton, F. (1907). Vox populi. *Nature 7*, 450–451.

Goldenberg, A., Abruzzo, J. M., Huang, Z., Schöne, J., Bailey, D., Willer, R., ... & Gross, J. J. (2023). Homophily and acrophily as drivers of political segregation. *Nature Human Behaviour*, 7(2), 219-230.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2), 231-237.

Hong, L., & Page, S. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385-16389.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Jayles, B., Kim, H. R., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., ... & Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, 114(47), 12620-12625.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, 142, 191-204.

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, *1*(6), 345-357.

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 281(1784), 20133305.

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., ... & Couzin, I. D. (2018). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface*, 15(141), 20180130.

Karachiwalla, R., & Pinkow, F. (2021). Understanding crowdsourcing projects: A review on the key design elements of a crowdsourcing initiative. *Creativity and innovation management*, *30*(3), 563-584.

Keck, S., & Tang, W. (2020). Enhancing the wisdom of the crowd with cognitive-process diversity: The benefits of aggregating intuitive and analytical judgments. *Psychological Science*, 31(10), 1272-1282.

Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., ... & Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. *Science*, *355*(6327), 820-826.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31), 8777-8782.

Lee, M.D., Villarreal, M., & Montgomery, L.E. (2023). Debiasing people's estimates with cognitive models to improve crowd aggregation. *Poster presented at the Annual Meeting of the Society for Judgment and Decision Making, San Francisco, CA*. [https://drive.google.com/file/d/1m7XIehCumtkeIfUvfw4Hxc-Yy8QOhuHS/view?usp=drive_link].

Lee, M.D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9, 259-273.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, 108(22), 9020-9025.

Madirolas, G., & de Polavieja, G. G. (2015). Improving collective estimations using resistance to social influence. *PLoS computational biology*, *11*(11), e1004594.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, *107*(2), 276.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*(5), 1106-1115.

Mohammed, S., & Ringseis, E. (2001). Cognitive diversity and consensus in group decision making: The role of inputs, processes, and outcomes. *Organizational behavior and human decision processes*, 85(2), 310-335.

Mussweiler, T., & Strack, F. (2001). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition*, *19*(2), 145-160.

Navajas, J., Armand, O., Moran, R., Bahrami, B., & Deroy, O. (2022). Diversity of opinions promotes herding in uncertain crowds. *Royal Society Open Science*, 9(6), 191497.

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2(2), 126-132.

Nobre, D. A., & Fontanari, J. F. (2020). Prediction diversity and selective attention in the wisdom of crowds. *arXiv preprint* arXiv:2001.10039.

Page, S. (2007). Making the difference: Applying a logic of diversity. *Academy of Management Perspectives*, *21*(4), 6-20.

Page, S. (2008). The difference. In *The Difference*. Princeton University Press.

Pescetelli, N., Rutherford, A., & Rahwan, I. (2021). Modularity and composite diversity affect the collective gathering of information online. *Nature Communications*, 12(1), 3195.

Ray, R. (2006). Prediction markets and the financial" wisdom of crowds". *The Journal of Behavioral Finance*, *7*(1), 2-4.

Röseler, L., Weber, L., Helgerth, K., Stich, E., Günther, M., Tegethoff, P., ... & Schütz, A. (2022). The Open Anchoring Quest Dataset: Anchored Estimates from 96 Studies on Anchoring Effects. *Journal of Open Psychology Data*, *10*(1), 16.

Ruggeri, K., Panin, A., Vdovic, M., Većkalov, B., Abdul-Salaam, N., Achterberg, J., ... & Toscano, F. (2022). The globalizability of temporal discounting. *Nature Human Behaviour*, 6(10), 1386-1397.

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature human behaviour*, 3(4), 329-336.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior research methods*, 42, 1096-1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24-31.

Sulik, J., Bahrami, B., & Deroy, O. (2022). The diversity gap: when diversity matters for knowledge. *Perspectives on Psychological Science*, 17(3), 752-767.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

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

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7), 645-647.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, *37*(1), 62-69.

Zimmerman, F., Garbulsky, G., Ariely, D., Sigman, M., & Navajas, J. (2022). Political coherence and certainty as drivers of interpersonal liking over and above similarity. *Science Advances*, 8(6), eabk1909.

Supplementary Information

Diversity Prediction Theorem

The diversity prediction theorem (Page, 2007) allows us to express the collective error of a crowd (E, defined as the squared difference between the crowd's mean answer and the correct answer) as two separate components: the predictive diversity (δ) defined as the squared difference between each individual answer A and the mean answer); and the mean individual error (ϵ), where individual error is defined as the squared difference between the individual answers and the correct answer). The proof of this mathematical identity is straightforward:

As expressed, we require the following definitions:

$$\boldsymbol{\varepsilon = \frac{1}{N}\sum_{q,i}\left(A^{(i)}{}_q - \theta_q\right)^2 = \frac{1}{N}\sum_{q,i}(A^{(i)}{}_q{}^2 + \theta_q{}^2 - 2\,.\,\theta_q\,.\,A^{(i)}{}_q),} \quad \text{[S1]}$$

$$\boldsymbol{\delta = \frac{1}{N}\sum_{q,i}\left(A^{(i)}{}_q - \mu_q\right)^2 = \frac{1}{N}\sum_{q,i}(A^{(i)}{}_q{}^2 + \mu_q{}^2 - 2\,.\,\mu_q\,.\,A^{(i)}{}_q),} \quad \text{[S2]}$$

$$\boldsymbol{E = \sum_q\left(\mu_q - \theta_q\right)^2} \quad \text{[S3]}$$

with N being the size of the population, the index *i* covering the N individuals, the index q covering all questions asked, θ_q being the correct answer for each question, $A^{(i)}{}_q$ being the answer provided by each individual *i* in response to question *q*, and $\mu_q = \frac{1}{N}\sum_i A^{(i)}{}_q$ being the mean answer for question *q*.

When we subtract ϵ and δ, the term ${A^{(i)}}_q{}^2$ cancels out. Since the parameters μ_q and θ_q are independent of i, we get:

$$\boldsymbol{\epsilon - \delta = \frac{1}{N}\sum\nolimits_{q,i}({\theta_q}^2 - {\mu_q}^2 - 2\,.\theta_q\,.{A^{(i)}}_q + 2\,.\mu_q\,.{A^{(i)}}_q) =} \qquad [S4]$$

$$\boldsymbol{= \sum_q ({\theta_q}^2 - {\mu_q}^2) + \frac{2}{N}\sum_{q,i}(\mu_q\,.{A^{(i)}}_q - \ \theta_q\,.{A^{(i)}}_q) =}$$

$$\boldsymbol{= \sum_q ({\theta_q}^2 - {\mu_q}^2 + \ 2.{\mu_q}^2 - 2\,\theta_q\,.\,\mu_q) =}$$

$$\boldsymbol{= \sum_q (\mu_q - \theta_q)^2 = E}$$

This identity indicates that the collective error $\boldsymbol{E}$ is exactly equal to the mean individual error minus the predictive diversity. One implication of this result is that, in principle, collective error can be reduced not only by decreasing the mean individual error (in other words, employing or selecting wiser individuals), but by increasing the predictive diversity of the crowd, which is related to its variance. It is important to note that this mathematical identity applies to convex error measures (relevant for all the problems addressed in this work) but fails in the case of concave error measures.

Conditions to Outperform the Wisdom of Crowds

To find the range of parameters under which the collective error of extremized crowds is lower than the non-extremized wisdom of crowds, the following expression should hold:

$$|\mu_x - \theta| < |\mu - \theta| \qquad [S5]$$

where μ_x is the mean of the extremized crowds, μ is the mean of the wisdom of crowds, and θ is the correct answer. Assuming that μ ≠ θ (i.e., the wisdom of crowds does not reach the correct answer, meaning that there is room for improvement), we have

$$\frac{|\mu_x - \theta|}{|\mu - \theta|} < 1 \qquad [S6]$$

Furthermore, without loss of generality, we assume that the crowd underestimated the correct answer, and so $\mu - \theta > 0$. (The procedure is analogous in the case of overestimation). We then expand the absolute value $\left|\frac{\mu_x - \theta}{\mu - \theta}\right|$, and multiply by $\mu - \theta$ on each side, reaching

$$-(\mu - \theta) < \mu_x - \theta < \mu - \theta \qquad [S7]$$

We then subtract $\mu - \theta$ on all sides, and reach

$$-2(\mu - \theta) < \mu_x - \mu < 0 \qquad [S8]$$

For the sake of clarity, we now simplify the problem and assume a constant anchoring weight w_0. This particular case is informative, as a constant anchoring weight would produce conditions which are less favorable for the model, given that it would lack access to the correct value θ through Eq. [4] (the proof for the general case can be found below). If we have a constant anchoring weight, using Eq. [2] and [3], and replacing them in Eq. [5], we have

$$\mu_x = w_0 \, \frac{A_L + A_H}{2} + (1 - w) \, \mu \qquad [S9]$$

If we define $\bar{A} = \frac{A_L + A_H}{2}$ (mean value of the anchors), and replace μ_x in Eq. [S8], we get

$$-2(\mu-\theta) < w_0\,\bar{A} + (1-w)\,\mu - \mu < 0 \qquad \text{[S10]}$$

Now we divide by w_0 and sum μ on all sides, reaching

$$-\frac{2}{w_0}(\mu-\theta) + \mu < \bar{A} < \mu \qquad \text{[S11]}$$

The inequality in Eq. [S11] implies that the range of values of $\bar{A}$ to outperform the wisdom of crowds is of the form:

$$\Delta_0 = \frac{2|\mu-\theta|}{w_0} \qquad \text{[S12]}$$

This result indicates that the conditions under which the model improves the wisdom of crowds depend on the midpoint of the anchors (its average $\bar{A}$). It also implies that the range of values for $\bar{A}$ that fulfills those conditions is at least twice the size of its error.

If we drop the assumption of constant anchoring weight, we can still follow the same procedure and reach, instead of Eq. [S11], the general expression

$$-\frac{4(\mu-\theta)}{(w_L+w_H)} + \mu + \frac{(w_L-w_H)}{(w_L+w_H)}\Delta A < \bar{A} < \mu + \frac{(w_L-w_H)}{(w_L+w_H)}\Delta A \qquad \text{[S13]}$$

This expression implies that the range of values where extremized crowds outperform the wisdom of crowds is given by Eq. [6].

Competition between Increasing Diversity and Accuracy

Given the predictions of the model, and the empirical results of the four behavioral experiments, one may ask if it is possible to increase predictive diversity (via anchoring) while, at the same time, reducing mean individual error. In principle, there is nothing in the Diversity Prediction Theorem that would prevent us from approaching both simultaneously (i.e., one could reduce the first term ε and increase the second term δ). However, in all the experiments and simulations presented here, we observed that every time we reduced collective error by increasing diversity, we also increased mean individual error. To better understand this phenomenon, we performed a new model-based analysis: we ran a series of simulations of the proposed model, while systematically varying the distance between the two anchors. Each of these simulations produced a set of estimates with roughly the same collective error, but different predictive diversity and mean individual error. Across simulations, we found a positive association between those two variables (Fig. S1, Pearson correlation coefficient, $r=0.998$, $p=2\text{x}10^{-62}$), suggesting that our approach establishes a tradeoff between increasing accuracy and diversity, and that they are indeed in competition.

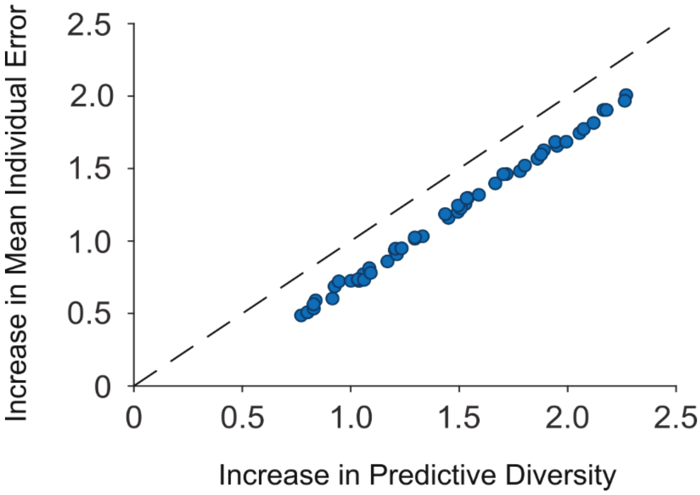

Fig. S1. Competition between increasing diversity and accuracy. Model simulations of changes in Mean Individual Error as a function of changes in Predictive Diversity. Model parameters were: $w_0 = 0.2$ (weight when the anchor is the correct answer), $A = -1$ (average of anchors), $\mu = 0.2$ (mean crowd answer), $n = 10{,}000$ (crowd size), $\sigma = 1$ (standard deviation of crowd answers), $\theta = 0$ (correct answer). Across simulations, we systematically varied the distance between anchors from $\Delta = 5$ to $\Delta = 10$ in steps of 0.1. Blue dots show results from different model specifications, and the dashed line depicts the identity.

Evaluation of Different Aggregation Methods

When analyzing the data of Experiment 2, we found strong evidence consistent with the existence of a bias towards intermediate probabilities (Figure S2). This holds true for both the non-extremized wisdom of crowds and the wisdom of extremized crowds.

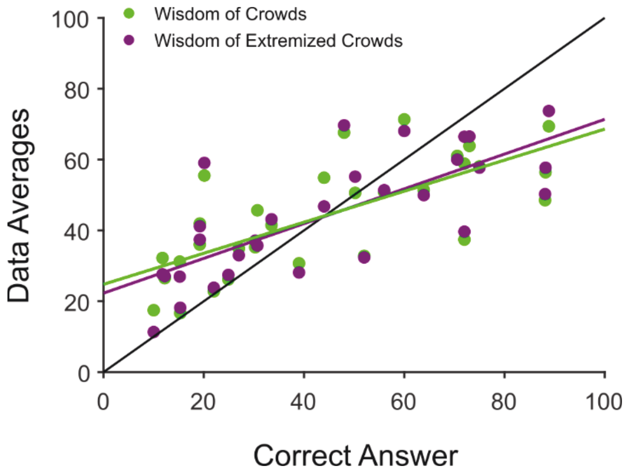

Fig. S2. Average Responses vs. Correct Answer. Average answers to each question as a function of the correct answer, for both the non-extremized wisdom of crowds (green dots) and the wisdom of extremized crowds (purple dots). Lines depict the best linear fits.

This result implies that, in principle, there is still room for improvement over our method by performing different aggregation procedures. We tested three different strategies. The first two strategies are the mean and the median, which have been widely used and compared since the first implementation of the wisdom of the crowds (Galton, 1907). The third aggregation strategy is a performance-weighted average (e.g., Collins et al, 2023). We estimated each participant's weight using a "leave-one-out" procedure. We first computed the mean individual error of each participant based on the responses to all but one question, and then normalized (so that they add up to 1) to define each participant's weight for the question that was left out.

As seen in Figure S3, in all cases, our method proves useful for reducing the error of the non-extremized wisdom of crowds (green). However, there are some differences between the aggregation methods. Using median instead of mean slightly increased the error of the non-extremized wisdom of crowds (specifically, for the largest crowd size N=50, unpaired t-test: t(998)=2.37, p=0.02; effect size: Cohen's d = 0.16 ± 0.09, 95% CL), but reduced the error of the extremized crowds (for the largest crowd size N=50, unpaired t-test: t(998)=6.38, $p=3x10^{-10}$; effect size: Cohen's d = 0.35 ± 0.09, 95% CL), resulting in the biggest difference between the two crowds. Performance-weighted averages worked best for smaller crowds, but show no statistical differences with the mean for the largest crowd size (unpaired t-test: t(998)=0.33, p=0.74; effect size: Cohen's d = 0.02 ± 0.09, 95% CL). Overall, these results further support not only the robustness but also the versatility of our method, which can be used with different aggregation strategies.

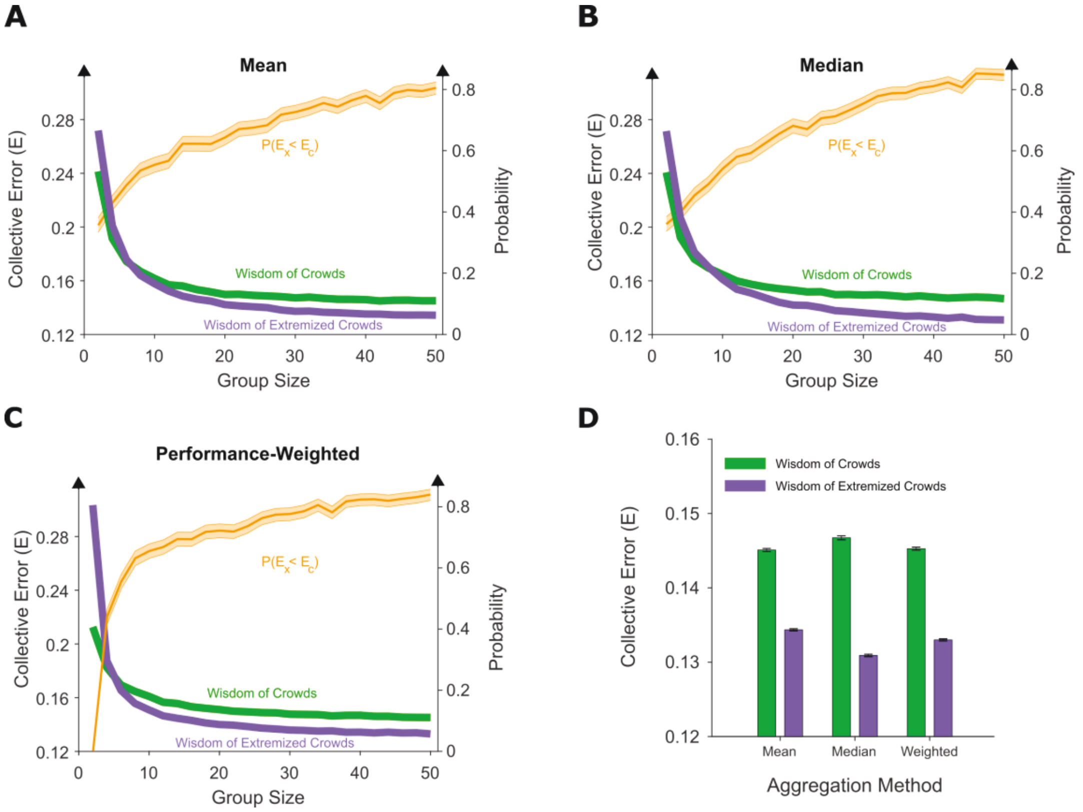

Fig. S3. Comparison of aggregation procedures for Experiment 2. (A-C) Collective error as a function of crowd size, for the non-extremized wisdom of crowds (green) and the extremized crowds (purple). The standard error of the curves is within the line width. The orange line corresponds to the probability of a sample from the distribution of collective errors of the extremized crowd (purple distribution, E_x) of being smaller than a sample from the distribution of collective errors of the normal crowd (green distribution, E_c). (A) Aggregation method: mean. (B) Aggregation method: median. (C) Aggregation method: performance-weighted mean. (D) Comparison of the collective error of the largest crowd (N=70) for each aggregation method. The wisdom of crowds is presented in green, and the wisdom of extremized crowds is presented in purple.

Self-generated Anchors and the Wisdom of the Inner Crowd

While Experiment 4 proved that the proposed method can be effective, even without prior knowledge by the experimenters, by asking participants to provide their own anchors, one issue remain unaddressed: could the wisdom-of-the-inner-crowd effect (Herzog & Hertwig, 2009) explain the efficacy of the method? If so, it could be argued that it is not the anchoring procedure, but the latter effect, the driver of error reduction in the extremized crowd for this Experiment. In order to assess this, and as was mentioned in the Methods section of Experiment 4, we asked participants in the control condition to provide new (different) estimates for each question. We then compared the wisdom of crowds, the wisdom of extremized crowds, and the wisdom of the inner crowd for each possible crowd size (Figure S4).

Results show that the wisdom of the inner crowd (which, according with Vul & Pashler (2008), gives a 10% improvement over the original estimates) is not statistically distinct from the wisdom of crowds in this experiment (for the largest crowd size N=70, unpaired t-test: t(998)=0.84, p=0.40; effect size: Cohen's d = 0.04 ± 0.09, 95% CL; Bayes factor for the null hypothesis: 18.01). This could have been expected, since assuming a 10% average improvement over the original individual estimates, for a large crowd involving estimates from different participants, the improvement becomes negligible when compared with the variance of the estimates themselves. Overall, this results show that our method's efficacy does not hinge on the wisdom-of-the-inner-crowd effect, but it is itself a distinct phenomenon.

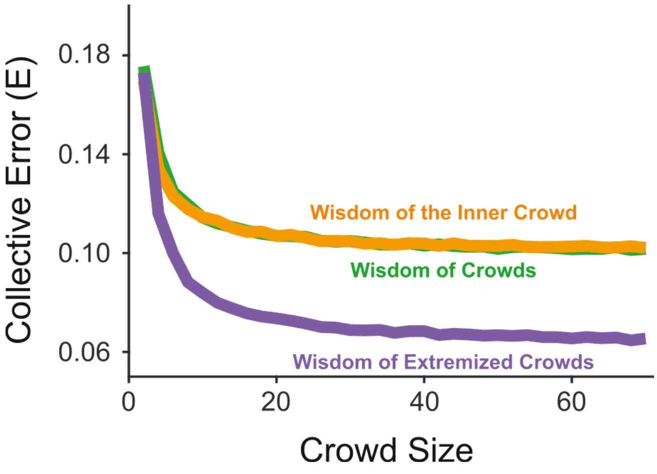

Fig. S4. Comparison with the Wisdom of the Inner Crowd for Experiment 4. Collective error as a function of crowd size, for the non-extremized wisdom of crowds (green), the wisdom of extremized crowds (purple), and the wisdom of the inner crowd (orange). The standard error of the curves is within the line width.

Definition of "Anchoring Index"

From Eq. [2] and [3], it follows that the anchoring index (Jacowitz & Kahneman, 1995) is defined as

$$w_j = \frac{\mu_j - \mu}{A_j - \mu} \qquad [16]$$

where *j* represents either *L* for the low anchor or *H* for the high anchor.

Experimental Design Variables

Experiment	N	Number of Questions	Type of Questions	Range of Answers	Country	Representative Sample	Time Limit	Incentive for Accuracy	Pre-reg
1	120	14	Estimation	[0,+inf)	USA	No	Yes	Yes	No
2	396	30	Estimation	[0,100]	Argentina	No	Yes	No	Yes
3	620	2	Forecasting	[0,+inf)	USA	Yes	No	Yes	Yes

Table S1: main design variables of Experiments 1-3.

Experiment 1: Questions and correct answers

Question	Correct Answer
What is the distance (miles) between Memphis (Tennessee, USA) and Oklahoma City (Oklahoma, USA)?	421.51
What is the distance (miles) between Milwaukee (Wisconsin, USA) and Faith (South Dakota, USA)?	714.18
What is the distance (miles) between Davenport (Iowa, USA) and Indianapolis (Indiana, USA)?	261.47
What is the distance (miles) between Paris (France) and Florence (Italy)?	550.81
What is the distance (miles) between Athens (Greece) and Rome (Italy)?	652.97
What is the height (to the tip, in yards) of the Eiffel Tower?	354.331
What is the height (to the tip, in yards) of Mount Vesuvius?	1400.92
How many floors does the Franklin Center have?	60
How many times does the word Jesus appear in the New Testament (New International Version, case insensitive)?	1273
How many times does the word Allah appear in the Qur`an (case insensitive)?	2699
How many times does the word Wisdom appear in the Qur`an (case insensitive)?	50
How many emperors did the Roman Empire have (Unified Empire only)?	71
How many heart transplantations were made in 2016 in the USA?	3191
How many bridges are there in Paris, France?	37

Table S2: Questions tested in Experiment 1 and correct answers

Experiment 2: Questions and correct answers

Question	Correct Answer
What percentage of the population of Argentina is not affiliated with any religion?	12.2
What percentage of the population of Argentina aged 20 or older is either overweight or obese?	52.0
What percentage of the members of the House of Representatives of Argentina are women?	39.0
What percentage of the population of Argentina is 14 years old or younger?	24.9
According to a representative survey conducted in 2013, what percentage of the population of Argentina believes that homosexuality is not morally acceptable?	27.0
According to a representative survey conducted in 2013, what percentage of the population of Argentina believes that abortion is not morally acceptable?	56.0
According to a representative survey conducted in 2014, what percentage of the population of Argentina believes they have good or very good health?	75.0
What percentage of male deaths between 15 and 24 years of age in Argentina were due to suicide?	15.3
What percentage of the population of Argentina aged 13 or older has a Facebook account?	60.0
According to a representative survey conducted in 2017, what percentage of the population of Argentina believes they are able to distinguish between real news and "fake news" (completely invented stories or facts)?	72.0
According to the 2010 census in Argentina, what percentage of the population between 3 and 18 years old attends an educational institution?	88.9
According to the 2010 census in Argentina, what percentage of the population lives in a housing unit without water discharge or without a toilet?	15.2
According to the 2010 census in Argentina, what percentage of women over 14 years old have never had a child?	30.3
According to the 2010 census in Argentina, what percentage of the population over 20 years old either has or is looking for a job?	70.6

According to the 2010 census in Argentina, what percentage of the population living in private homes has health coverage?	63.9
What percentage of women aged between 18 and 60 years old have jobs in Argentina?	50.2
What percentage of the Argentine population has access to the internet at home, either through a computer or mobile device?	73.0
According to a representative survey conducted in 2013, what percentage of the Argentine population believes that it is not morally acceptable for unmarried adults to have sexual relationships?	22.0
What percentage of the Argentine population lives in a housing unit that they own?	72.0
What percentage of the Argentine population owns a smartphone?	48.0
According to a representative survey from 2013, what percentage of the Argentine population believes that most people are trustworthy?	19.2
What percentage of the Argentine population believes that some vaccines can cause autism in healthy children?	10.0
In Argentina, what percentage of the total real estate wealth belongs to the richest 1% of the population?	44.0
According to the 2010 census in Argentina, what percentage of people over 20 years old have completed high school?	20.1
According to the 2010 census in Argentina, what percentage of the population over 80 years old was born in another country?	11.8
According to a representative survey conducted in 2013, what percentage of the population of Argentina is not at all interested in politics?	30.7
According to the 2010 census in Argentina, what percentage of the population living in the Santa Fe province is under 65 years old?	88.2
According to the 2010 census in Argentina, what percentage of the population is single?	33.5
According to the 2010 census in Argentina, what percentage of the population does not live in an apartment?	88.1
According to the 2010 census in Argentina, what percentage of people aged 20 to 29 both work and study?	19.2

Table S3: Questions tested in Experiment 2 and correct answers

Experiment 3: Questions and correct answers

Question	Correct Answer
How many new deaths by COVID-19 do you think there will be in the United States in the upcoming week (27 July – 2 August)?	7987
How many new cases by COVID-19 do you think there will be in the United States in the upcoming week (27 July – 2 August)?	442,417

Table S4: Questions tested in Experiment 3 and correct answers

Experiment 4: Questions and correct answers

Question	Correct Answer
What is the height of the “Obelisco de Buenos Aires”?	67
What is the height of the “Catedral de La Plata”?	112
What is the height of the “Palacio de Aguas Corrientes”?	21
What is the height of the “Palacio del Congreso de la Nación Argentina”?	80
What is the height of the “Torre de los Ingleses de Retiro”?	60
What is the height of the Eiffel Tower (to the tip)?	330
What is the height of the Empire State Building (to the tip)?	443
What is the height of the Buckingham Palace?	24
What is the height of the Tower of Shanghai?	632
What is the height of the Tower of Pisa?	56

Table S5: Questions tested in Experiment 4 and correct answers