Taking a Disagreeing Perspective Improves the Accuracy of People's Quantitative Estimates

Please check the document version of this publication:

Research Article

Taking a Disagreeing Perspective Improves the Accuracy of People's Quantitative Estimates

Philippe P. F. M. Van de Calseyde[1] [iD] and Emir Efendić[2] [iD]

[1]Human Performance Management Group, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, and [2]Marketing and Supply-Chain Management, School of Business and Economics, Maastricht University

Abstract

Many decisions rest on people's ability to make estimates of unknown quantities. In these judgments, the aggregate estimate of a crowd of individuals is often more accurate than most individual estimates. Remarkably, similar principles apply when multiple estimates from the same person are aggregated, and a key challenge is to identify strategies that improve the accuracy of people's aggregate estimates. Here, we present the following strategy: Combine people's first estimate with their second estimate, made from the perspective of someone they often disagree with. In five preregistered experiments ($N = 6,425$ adults; $N = 53,086$ estimates) with populations from the United States and United Kingdom, we found that such a strategy produced accurate estimates (compared with situations in which people made a second guess or when second estimates were made from the perspective of someone they often agree with). These results suggest that disagreement, often highlighted for its negative impact, is a powerful tool in producing accurate judgments.

People often make estimates of some unknown quantities or events. In these types of judgments, a well-known phenomenon is that the average estimate of a crowd of individuals is often more accurate than most individual estimates (Galton, 1907; Surowiecki, 2005), and crowds have been used to improve judgments in areas such as economic forecasts (Clemen, 1989), medical diagnoses (Kurvers et al., 2016), weather forecasting (Baars & Mass, 2005), and scientific research (Altmejd et al., 2019; Gordon et al., 2021).

The "wisdom of crowds" arises from the (mathematical) principle whereby aggregating multiple imperfect, yet diverse, estimates diminishes the role of errors (Herzog & Hertwig, 2009; Stroop, 1932). That is, when multiple estimates are sufficiently diverse and independent, averaging increases accuracy by canceling out errors across individuals (Vul & Pashler, 2008). However, although this approach is beneficial, it is often not feasible for a single person to collect the estimates of multiple individuals. Remarkably, research suggests that the same principles underlying the wisdom of crowds also apply when multiple estimates from the same person are aggregated—a phenomenon known as the "wisdom of the inner crowd" (Herzog & Hertwig, 2009; Van Dolder & Van den Assem, 2018; Vul & Pashler, 2008).

It is not clear from the outset why aggregating multiple estimates from the same person would be beneficial. If a person's first estimate represents their best guess, then any other estimate would simply add noise

Corresponding Author:
Philippe P. F. M. Van de Calseyde, Eindhoven University of Technology, Department of Industrial Engineering and Innovation Sciences, Human Performance Management Group
Email: P.P.F.M.v.d.Calseyde@tue.nl

(Hourihan & Benjamin, 2010; Vul & Pashler, 2008). An alternative account based on probabilistic representations, however, posits that averaging estimates from the same person cancels out the errors that permeate people's judgments. According to this account, people's initial estimates represent samples drawn from an internal distribution of possible estimates, where second estimates are resampled guesses from that same distribution (Vul & Pashler, 2008; Wallsten et al., 1997). When second, resampled estimates are sufficiently diverse, averaging increases accuracy by canceling out errors across estimates (Ariely et al., 2000; Herzog & Hertwig, 2009; Keck & Tang, 2020; Litvinova et al., 2020).

With such a powerful tool available to individuals, a key challenge is to identify strategies that can help improve the accuracy of people's aggregate estimates (Herzog & Hertwig, 2014a). Research so far agrees that the inner crowd falters when people anchor too heavily on their first estimate when generating a second guess, thereby reducing diversity and independence (Herzog & Hertwig, 2014a; Vul & Pashler, 2008). At least two methods have been applied to negate this. The first relies on the passage of time. For example, the benefits of aggregation tend to be higher with the introduction of time delays between both estimates (Steegen et al., 2014; Van Dolder & Van den Assem, 2018; Vul & Pashler, 2008). In these cases, the passage of time effectively deanchors people from their first estimate (presumably because they forget their initial estimate), thereby improving the diversity and independence of both estimates. A second method to increase diversity and independence is to rely on the mind's ability to construct alternative, opposing realities (Herzog & Hertwig, 2009, 2014a, 2014b). A demonstrated way to do this has been through "dialectical bootstrapping," in which people are prompted to base their second estimate on different assumptions and considerations (Herzog & Hertwig, 2009). These dialectical estimates ideally result in errors with different signs relative to first estimates, and there are different techniques to elicit such dialectical estimates. One technique, based on the "consider-the-opposite" strategy (Lord et al., 1984), instructs people to actively question the accuracy of their first estimate when generating a second guess. This technique has been shown to increase the accuracy of people's aggregate estimate by getting the same person to generate first and second estimates that are more diverse and independent (Herzog & Hertwig, 2009, 2014b). In the present research, we similarly relied on the mind's ability to construct opposing realities by prompting people to complement their initial estimate with a second estimate made from the perspective of someone they often disagree with.

Statement of Relevance

In today's polarized society, disagreement is associated with conflict and division, but are there also benefits to disagreement? By utilizing people's ability to take the perspective of others, we propose that disagreement is a powerful tool for producing accurate estimates. In five experiments, people made estimates of unknown quantities from various perspectives. Following principles of within-person aggregation, we found that aggregating people's first estimate with their second estimate, made from the perspective of someone they often disagree with, produced accurate estimates. In explaining this accuracy, we found that taking a disagreeing perspective prompts people to consider estimates they normally would not consider to be viable options, resulting in first and second estimates that are more diverse and independent (and by extension more accurate when aggregated). Together, these results underscore the importance of perspective taking and disagreement as strategies to improve the accuracy of people's quantitative estimates.

Perspective taking refers to people's ability to consider situations and events from the viewpoint of others (Piaget, 1932/1965). It has been associated with many positive outcomes, such as altruistic behaviors, decreased stereotype expressions, and increased creativity (Batson et al., 1997; Galinsky & Moskowitz, 2000; Hoever et al., 2012). However, according to the principles of within-person aggregation, simply getting people to take the perspective of others would not be enough. What is needed is to add an estimate from the perspective of someone whose views are substantially different—in other words, to create a diverse inner crowd. To do this, we suggest using an oft-encountered component of people's interaction with others—disagreement. More specifically, as a viable method to obtain more diverse estimates, we propose to combine people's initial estimate with their second estimate made from the perspective of someone they often disagree with.

Disagreement is often decried as an undesirable component of people's interactions with others. In today's polarized society, disagreement has been associated with conflict, division, and misinformation (Kennedy & Pronin, 2008; Reeder et al., 2005; Sunstein, 2002). However, although disagreement is generally undesirable, research in group decision-making indicates that it may

actually be beneficial when groups address complex problems, such as making estimates of unknown quantities or events (de Oliveira & Nisbett, 2018; Hong & Page, 2004; Mutz, 2006; Page, 2008). These effects occur because of the notion that disagreeing individuals tend to produce more diverse estimates, and by extension errors, which are canceled out across group members when averaged (Page, 2008). It is precisely this aspect of disagreement that we rely on in our pursuit to foster more diverse estimates from the same individual. More specifically, we surmise that just as disagreement between different individuals is beneficial for the wisdom of the crowds, so too, through perspective taking, will this be beneficial for the wisdom of the inner crowd.

To understand the benefits of disagreement, we tested the hypothesis (in Experiment 3) that taking a disagreeing perspective leads to two distinct observations. First, from a disagreeing perspective, people are more likely to consider estimates that are strikingly different from their own guesses, thereby opening the sampling space of possible second estimates. And second, people are more likely to adopt these different estimates as their second estimates when viewing problems from a disagreeing perspective, leading to first and second estimates that are more diverse and independent. These conjectures follow from earlier work on anchoring showing that people typically avoid making second estimates that are strikingly different from prior estimates or anchors (Epley & Gilovich, 2006; Lewis et al., 2019; Tversky & Kahneman, 1992). Making an estimate from a disagreeing perspective is expected to attenuate this tendency, given that disagreeing others (almost by default) consider and adopt entirely different estimates as one's own estimate. However, although taking a disagreeing perspective is generally beneficial, the final experiment (Experiment 4) identified a situation in which taking a disagreeing perspective backfired, undermining the benefit of averaging (i.e., in situations in which second estimates were likely to be made in the wrong direction).

The Present Research

For easier reading, we first present general methodological information that concerns all five experiments. Ethical approval for all experiments was obtained from the ethical review board at Eindhoven University of Technology (Reference No. ERB2020IEIS29). For all experiments, we report how we determined the sample size, all data exclusions (if any), all manipulations, and all measures. The questions used in all experiments can be found in the Supplemental Material available online. Data, code, and materials are publicly available on OSF at https://osf.io/qsxp8/. All experiments' hypotheses,

designs, and main analyses were preregistered[2] (see the Open Practices section for links).

Sample-size estimation for all experiments was based on a priori power analyses using G*Power (Version 3.1.9.4; Faul et al., 2007). For Experiments 1a, 1b, 2, and 3, analyses determined that 410 participants per condition would be necessary to achieve a Cohen's *d* effect size of 0.30 with 99% power and that 394 participants per condition would be necessary to achieve a Cohen's *d* effect size of 0.20 with more than 80% power. For Experiment 4, analyses determined that 290 participants per condition would be necessary to achieve a Cohen's *d* effect size of 0.30 with 95% power and that 253 participants per condition would be necessary to achieve a Cohen's *d* effect size of 0.25 with 80% power. Alpha was set at .05. For all experiments, we stopped data collection when we reached the predetermined sample size. Following previous studies on the inner crowd, to verify the accuracy of people's estimates, we relied on the mean square error[3] obtained by squaring the subtraction of the true answers from the estimations and then averaging.

For analysis, we used mixed-effect models, which allowed us to make more generalizable claims across a wide range of participants and questions by employing random intercepts for participants and questions (Judd et al., 2012). We fitted the models using *lme4* (Version 1.1-27.1; Bates et al., 2015) and produced *p* values using the Satterthwaite approximations for degrees of freedom from *lmerTest* (Version 3.1-3; Kuznetsova et al., 2017). Because there is little agreement on how to calculate effect sizes for mixed models, we report classic Cohen's *d* or d_z effects calculated from the *t* values of the fixed-effect results obtained in the models (Cohen, 1988). For comparison of correlation coefficients between experimental conditions, we took a two-step approach. Because participants responded to multiple questions twice, we first calculated the correlation between the errors of first and second estimates (i.e., the true answer subtracted from an estimate[1]) for each participant. We then compared these Pearson's *r* values across experimental conditions using independent-samples *t* tests and calculated Cohen's *d* effect sizes.

Experiments 1a and 1b

Method

In Experiment 1a, participants made two weight estimates of 10 objects shown in pictures (see Table S1 in the Supplemental Material). In Experiment 1b, we used a different estimation task: Participants made two estimates of six questions on a scale ranging from 0% to 100%. The questions' true answers were obtained from various online sources (e.g., Wikipedia for Experiment

1a and the *The World Factbook*, Central Intelligence Agency, 2020, 2021, for Experiment 1b). For Experiment 1a, we recruited 900 participants using Amazon Mechanical Turk (MTurk). Following the preregistration plans, we excluded participants who failed the instructional check and those who said they looked up the answers, leaving a final sample of 880 U.S. participants (age: *Mdn* = 36 years, interquartile range [IQR] = 16 years; 51% female). For Experiment 1b, we recruited 1,000 participants using Prolific. Excluding those who failed the instructional check and those who looked up the answers resulted in a final sample of 894 UK participants (age: *Mdn* = 35 years, IQR = 20 years; 69% female). After making their first estimate for all questions, half of the participants were told to make a second guess, and the other half were instructed to make their second estimate from the perspective of a friend they often disagree with.

Participants were instructed not to look up the true answers during the study. They were randomly presented with the questions in two estimation stages (histograms for the distribution of participants' answers on both estimates for all five experiments can be accessed at https://osf.io/q3tfh/). Participants were not told at the beginning of the experiment that they would be asked to make an additional, second estimate. In the first stage, participants simply provided their own estimates to the questions. The instructions for the second estimation stage were different depending on the condition. For the *self-perspective condition*, participants were told,

> We will now ask you to provide a second guess at the answer to each of the [ten/six] questions you were asked in the first session. These answers should not be the same as your previous answers: these should reflect your 'second guess'.

For the *disagreeing-perspective condition*, participants were told,

> Now picture a friend whose views and opinions are very different from yours. To illustrate, when discussing politics, you often find yourself disagreeing on various issues. How would he or she answer these [ten/six] questions? Please answer these questions again, but now as this friend.

After responding to the questions, participants were asked to provide their age and gender. In addition, they were presented with a manipulation-check question instructing them to choose a particular option in a multiple-choice array and a question asking them whether they looked up any of the answers to the questions.

Results

Correlations. Comparing the two experimental conditions, we found that our instructions led to lower correlations when second estimates were made from a disagreeing perspective (Experiment 1a: mean $r_{\text{disagreeing}}$ = .54 vs. mean r_{self} = .71; Experiment 1b: mean $r_{\text{disagreeing}}$ = .34 vs. mean r_{self} = .73). In both experiments, these two correlation coefficients were significantly different (Experiment 1a: d = 0.44, Experiment 1b: d = 0.98; both ps < .001), indicating that participants in the disagreeing-perspective condition produced more diverse estimates and errors compared with participants in the self-perspective condition (see Figs. 1a and 1b; scatterplots for each question separately for all five experiments can be accessed at https://osf.io/q3tfh/).

Inner-crowd effects. For the inner crowd to be more accurate, the aggregate of both estimates should have a lower error than a person's first estimate alone. Taking into account both conditions (overall) and looking at the self- and disagreeing-perspective conditions separately, we found an inner-crowd effect in Experiments 1a and 1b (see Table 1 for summary statistics). The average of both estimates had a lower mean square error than the first and second estimates alone, respectively (for descriptive statistics, see Tables S2a and S2b in the Supplemental Material).

Benefit of averaging. Would participants in the disagreeing-perspective condition benefit more from averaging their estimates than participants in the self-perspective condition? To test this, we calculated the benefit of averaging by subtracting the square error of average estimates from the square error of first estimates (similar procedures have been used before; Herzog & Hertwig, 2009; Steegen et al., 2014; Vul & Pashler, 2008). The higher this number, the larger the benefit of averaging (i.e., the more accurate the inner crowd). Overall, the results indicated that in both Experiments 1a and 1b[4] participants in the disagreeing-perspective condition indeed benefited more from averaging their estimates than participants in the self-perspective condition (Experiment 1a: d = 0.16, p = .02; Experiment 1b: d = 0.18, p = .01).

Bracketing. To more concretely test whether people in the disagreeing-perspective condition benefited more from averaging, we looked at bracketing rates across conditions. Bracketing is a key component underpinning the benefit of aggregating multiple estimates (Larrick & Soll, 2006). It refers to the observation that if two estimates are on the opposite sides of the true answer, thus bracketing it (i.e., one overestimating the true answer and the other underestimating it), aggregating them will typically result in a more accurate average estimate
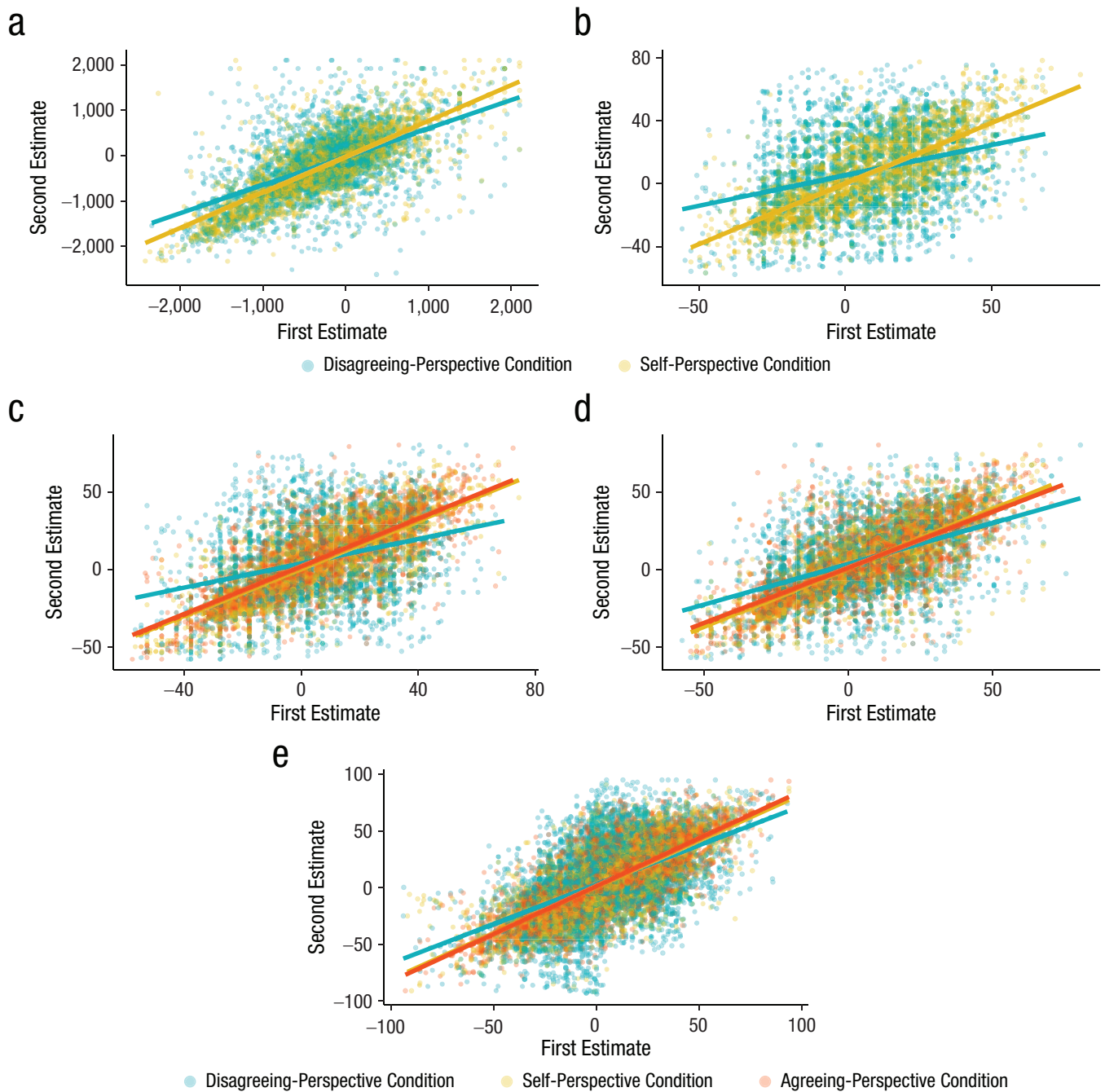
Fig. 1. Signed errors in Experiment 1a (a), Experiment 1b (b), Experiment 2 (c), Experiment 3 (d), and Experiment 4 (e). Each dot represents the correlation between the error of a participant's first estimate and the error of that participant's second estimate in response to a particular question, separately for the disagreeing-perspective condition (in all experiments), the self-perspective condition (in all experiments), and the agreeing-perspective condition (in Experiments 2–4). Solids lines are lines of best fit.

(Larrick & Soll, 2006; Soll & Larrick, 2009). Consequently, for each question, we verified whether the question's true answer was bracketed by the two estimates. As expected, the bracketing rate was much higher in the disagreeing-perspective condition at 29% (Experiment 1a) and 38% (Experiment 1b), compared with the self-perspective condition, in which 19% (Experiment 1a) and 20% (Experiment 1b) of people's estimates bracketed the

questions' true answers (Experiment 1a: $d = 0.56$, $p < .001$; Experiment 1b: $d = 0.80$, $p < .001$).

Experiment 2

People who made a second estimate from the perspective of a person they often disagree with benefited more from averaging than people who simply made a second

Table 1. Inner-Crowd Effect for Experiments 1a and 1b: Comparisons of the Average of Two Estimates With the First and Second Estimate

Experiment and comparison	Overall		Disagreeing-perspective condition		Self-perspective condition	
	d_z	p	d_z	p	d_z	p
Experiment 1a						
First estimate vs. average	0.19	< .001	0.23	< .001	0.15	.002
Second estimate vs. average	0.28	< .001	0.40	< .001	0.16	.001
Experiment 1b						
First estimate vs. average	0.22	< .001	0.28	< .001	0.16	< .001
Second estimate vs. average	0.53	< .001	0.79	< .001	0.24	< .001

guess. Experiment 2 provided an important extension. Specifically, we included a third experimental condition in which participants were instructed to take the perspective of a person they often agree with. We included this condition to underscore the need to take a disagreeing perspective to improve the accuracy of people's inner crowds.

Method

The procedure of this experiment was similar to that of Experiment 1b. However, we added an additional condition (i.e., the *agreeing-perspective condition*) in which the instructions for the second estimation stage were,

> Now picture a friend whose views and opinions are very similar to yours. To illustrate, when discussing politics, you often find yourself agreeing on various issues. How would he or she answer these six questions? Please answer these questions again, but now as this friend.

We recruited 1,425 participants using MTurk. After excluding those who failed the instructional check and those who said they looked up the answers, we obtained a final sample of 1,389 U.S. participants (age: *Mdn* = 35 years, IQR = 16 years; 44% female).

Results

Correlations. The estimates' errors were highly correlated in the self-perspective and the agreeing-perspective conditions (mean r_{self} = .73 vs. mean $r_{agreeing}$ = .74). This correlation was much lower in the disagreeing-perspective condition (mean $r_{disagreeing}$ = .32; see Fig. 1c). Comparing these correlation coefficients, we found that participants in the disagreeing-perspective condition produced more diverse errors than participants in both the self-perspective condition ($d = 0.99$) and agreeing-perspective condition

($d = 1.00$; both ps < .001), whereas there was no difference in the diversity of errors between the self- and agreeing-perspective conditions ($d = 0.02$, $p = .81$).

Inner-crowd effects. There was again an inner-crowd effect overall (i.e., across all three conditions) and in the three conditions separately (see Table 2; for descriptive statistics, see Table S3 in the Supplemental Material).

Benefit of averaging. There was no difference between the self- and agreeing-perspective conditions with regard to benefit of averaging ($d = 0.03$, $p = .61$). Importantly, participants in the disagreeing-perspective condition again benefited more from averaging both estimates, compared with the self- and agreeing-perspective conditions ($d = 0.18$, $p = .01$, and $d = 0.21$, $p = .001$, respectively).

Bracketing. With 21% and 20% of people's estimates bracketing the questions' true answers, there was no difference in bracketing rates between the self- and agreeing-perspective conditions ($d = 0.08$, $p = .25$). Crucially, however, the bracketing rate was again higher in the disagreeing-perspective condition: 39% of people's estimates bracketed the questions' true answers, compared with both the self-perspective ($d = 0.85$, $p < .001$) and agreeing-perspective ($d = 0.90$, $p < .001$) conditions.

Experiment 3

In Experiment 3, we tested the proposed mechanism explaining our observation of more diversity and independence when second estimates are made from a disagreeing perspective. Earlier work on inner crowds suggests that people typically anchor too heavily on first estimates when generating a second guess, thereby not producing diverse enough estimates and errors (Herzog & Hertwig, 2009; Vul & Pashler, 2008). Making an estimate from a disagreeing perspective was expected to attenuate this tendency, given that disagreeing others

Table 2. Inner-Crowd Effects for Experiments 2 and 3: Comparisons of the Average of Two Estimates With the First and Second Estimate

Experiment and comparison	Overall		Disagreeing-perspective condition		Self-perspective condition		Agreeing-perspective condition	
	d_z	p	d_z	p	d_z	p	d_z	p
Experiment 2								
First estimate vs. average	0.19	< .001	0.28	< .001	0.15	.001	0.14	.003
Second estimate vs. average	0.44	< .001	0.80	< .001	0.25	< .001	0.23	< .001
Experiment 3								
First estimate vs. average	0.21	< .001	0.25	< .001	0.17	< .001	0.23	< .001
Second estimate vs. average	0.33	< .001	0.56	< .001	0.22	< .001	0.19	< .001

(almost by default) consider and adopt entirely different estimates as one's own estimate.

Method

The design and procedure was similar to that of Experiment 2. However, before making their second estimate, participants in the self-perspective condition were asked, "What is the most extreme estimate (either extremely high or extremely low) that you would consider as second guess to this question?" In the agreeing- and disagreeing-perspective conditions, participants were asked, "What is the most extreme estimate (either extremely high or extremely low) that your friend would consider as answer to this question?" We recruited 1,500 participants using MTurk. After excluding those who failed the instructional check and those who said they looked up the answers, we obtained a final sample of 1,426 U.S. participants (age: Mdn = 36 years, IQR = 17 years; 48% female).

Results

Correlations. The estimates' errors were again highly correlated in the self- and agreeing-perspective conditions (mean r_{self} = .74 vs. mean $r_{agreeing}$ = .72). This correlation was much lower in the disagreeing-perspective condition (mean $r_{disagreeing}$ = .46; see Fig. 1d). Comparing these correlation coefficients, we found that participants in the disagreeing-perspective condition again produced more diverse errors than participants in both the self-perspective condition (d = 0.78) and agreeing-perspective condition (d = 0.71; both ps < .001), whereas there was no difference in error diversity between the self-perspective and agreeing-perspective conditions (d = 0.07, p = .26).

Inner-crowd effects. There was an inner-crowd effect overall (i.e., across all three conditions) and in the three

conditions separately (see Table 2; for descriptive statistics, see Table S4 in the Supplemental Material).

Benefit of averaging. Participants in the agreeing-perspective condition benefitted slightly more from averaging than did those in the self-perspective condition (d = 0.13, p = .04). Importantly, participants in the disagreeing-perspective condition again benefited more from averaging both estimates, compared with the self-perspective condition (d = 0.15, p = .02). However, there was no difference between the agreeing- and disagreeing-perspective conditions (d = 0.04, p = .55),[5] although the effect was in the right direction: The benefits of averaging were numerically higher in the disagreeing-perspective condition.

Bracketing. With 20% and 21% of people's estimates bracketing the questions' true answers, there was no difference in bracketing rates between the self- and agreeing-perspective conditions (d = 0.05, p = .42). Crucially, however, the bracketing rate was again higher in the disagreeing-perspective condition: 33% of people's estimates bracketed the questions' true answers, compared with both the self-perspective condition (d = 0.64, p < .001) and the agreeing-perspective condition (d = 0.61, p < .001).

Extreme-estimate analysis. To test the proposition that taking a disagreeing perspective prompts people to consider more extreme estimates as possible answers to a question, we computed the (absolute) difference score between each participant's first estimate and the most extreme estimate that they (or their friend) would consider as an answer. As expected, there was no difference between participants in the agreeing- and self-perspective conditions (d = 0.04, p = .51). Importantly however, participants in the disagreeing-perspective condition considered far more extreme estimates as possible answers than participants in either the self-perspective condition (d = 0.41, p < .001) or the agreeing-perspective condition (d = 0.46, p < .001). Moreover, to test whether participants in the disagreeing-perspective condition would also be more

inclined to adopt these extreme estimates as their second answers, we computed the (absolute) difference score between each participant's second estimate and the most extreme estimate. The lower this number, the closer the second estimate was to the most extreme estimate.

As expected, there was no difference between participants in the agreeing- and self-perspective conditions ($d = 0.12$, $p = .06$). Importantly, participants in the disagreeing-perspective condition made second estimates much closer to the extreme estimate than either the participants in the self-perspective condition ($d = 0.29$, $p < .001$) or the agreeing-perspective condition ($d = 0.14$, $p = .03$; for descriptive statistics, see Table S5 in the Supplemental Material). The willingness of participants to adopt these extreme estimates as answers is noteworthy, given people's general propensity to avoid making extreme judgments (Lewis et al., 2019). This aversion seems to dissipate when second estimates are made from the viewpoint of disagreeing others. Interestingly, even if people made second estimates equally close to their most extreme estimate from a disagreeing perspective, they would still produce more diverse estimates, given that these extreme estimates are generally more extreme. Overall, these results underscore the conjecture that taking a disagreeing perspective prompts people to consider and adopt second estimates that are strikingly different from their initial estimate, rendering a set of estimates that is more diverse and independent.

Experiment 4

The final experiment identified a situation in which taking a disagreeing perspective backfired. Specifically, this was expected in situations where a question's true answer lies close to the lower or upper end of a scale (e.g., if the true answer is 2% or 98% on a scale from 0% to 100%) and when a person's initial estimate is close to this answer. For example, imagine being asked the following question: "What percentage of China's population identifies as Christian?" The true answer to this question is 5.1%, and if you are like most people, your first estimate probably leaned toward the lower end of the scale (say your first estimate was 10%). Given the position of the question's true answer and your first estimate, your second estimate is likely (in general) to move away from the answer toward the opposite side of the scale (Juslin et al., 2000), effectively hurting the accuracy of your average estimate. Importantly, such a movement is expected to be especially harmful when second estimates are made from a disagreeing perspective because, given people's propensity to adopt more extreme estimates from such a perspective (see Experiment 3), these estimates move away from the true answer to a much greater extent (resulting in an average estimate that is far worse than the initial estimate).

Method

We gathered data in two waves. We preregistered our hypotheses and analysis plan for the second wave. Because the procedures in the two waves were identical, we decided to combine them (analyzing the data separately yielded similar results, which can be accessed at https://osf.io/ewpyq/). The procedure of this experiment was similar to that of Experiment 2 in all but two respects. First, we added an additional six questions to make 12 questions in total. Second, we categorized the questions according to where the true answer fell—that is, whether the true answer was in the middle of the scale or the end of the scale (0%–10% or 90%–100%). Participants thus made two estimates about a set of 12 questions, all of which had a true answer that was in the 0% to 100% range. Crucially, half of the questions' true answers were close to the lower or upper end of the scale, from 0% to 10% and 90% to 100%. For the other half of the questions, true answers were relatively far from the end of the scale (e.g., 58%). Combining the two data-wave collections, we recruited 1,889 participants using MTurk. As in the prior experiments, we excluded those who failed the instructional check and those who said they looked up the answers, leaving a final sample of 1,836 U.S. participants (age: *Mdn* = 36 years, IQR = 17 years; 51% female).

Results

Correlations. Correlations between the estimates' errors in the self- and agreeing-perspective conditions were again high (mean $r_{\text{self}} = .77$, mean $r_{\text{agreeing}} = .79$). This correlation was lower in the disagreeing-perspective condition (mean $r_{\text{disagreeing}} = .57$; see Fig. 1e). Participants in the disagreeing-perspective condition produced more diverse errors than participants in both the self-perspective ($d = 0.83$) or agreeing-perspective ($d = 0.96$) conditions (both ps < .001). The difference between the self- and agreeing-perspective conditions was also significant ($d = 0.15$, $p = .01$). Overall, the disagreeing-perspective condition again produced more diverse errors.

Inner-crowd effects. Taking into account both types of questions (mid-scale and end-scale questions) and all three conditions, we did not find an inner-crowd effect (see Table 3). Importantly, and in line with our proposal, results showed that the perspective-taking instructions had a markedly different impact when the mid-scale and end-scale questions were considered separately. For the mid-scale questions, there was an inner-crowd effect similar to those in the previous experiments. However, when looking at the end-scale questions for the disagreeing-perspective condition, we found that the average of both estimates had a much higher error than the first estimate

Table 3. Inner-Crowd Effect for Experiment 4: Comparisons of the Average of Two Estimates With the First and Second Estimate

Question type and comparison	Overall		Disagreeing-perspective condition		Self-perspective condition		Agreeing-perspective condition	
	d_z	p	d_z	p	d_z	p	d_z	p
Both question types								
First estimate vs. average	0.03	.16	0.05	.20	0.08	.05	0.08	.06
Second estimate vs. average	0.54	< .001	0.88	< .001	0.36	< .001	0.29	< .001
Mid-scale questions								
First estimate vs. average	0.19	< .001	0.28	< .001	0.14	.001	0.15	< .001
Second estimate vs. average	0.44	< .001	0.70	< .001	0.34	< .001	0.26	< .001
End-scale questions								
First estimate vs. average	−0.09	< .001	−0.26	< .001	0.01	.90	0.01	.71
Second estimate vs. average	0.38	< .001	0.66	< .001	0.21	< .001	0.19	< .001

alone (for descriptive statistics, see Table S6 in the Supplemental Material).

Benefit of averaging. For the mid-scale questions, we found no difference between the self- and agreeing-perspective conditions ($d = 0.05$, $p = .39$), whereas the benefit of averaging was again higher for participants in the disagreeing-perspective condition compared with the self-perspective ($d = 0.28$, $p < .001$) and agreeing-perspective ($d = 0.24$, $p < .001$) conditions. Thus, for the mid-scale questions, the results echo those obtained in the previous experiments. For the end-scale questions, there was no difference between the self- and agreeing-perspective conditions ($d = 0.05$, $p = .39$). However, in the disagreeing-perspective condition, averaging was actually much more disadvantageous than in the self-perspective ($d = −0.41$, $p < .001$) and agreeing-perspective ($d = −0.40$, $p < .001$) conditions.

Bracketing. For the mid-scale questions, with 23% and 21% of the estimates bracketing the questions' true answers, there was slightly more bracketing in the self-perspective than the agreeing-perspective condition ($d = 0.13$, $p = .03$). Importantly, as expected, the degree of bracketing was again higher in the disagreeing-perspective condition: 37% of the estimates bracketed the questions' true answers, compared with both the self-perspective ($d = 0.65$, $p < .001$) and agreeing-perspective ($d = 0.78$, $p < .001$) conditions.

Focusing on the end-scale questions, we generally saw lower rates of bracketing. With 13% and 11% of the estimates bracketing the questions' true answers, there was slightly more bracketing in the self-perspective than the agreeing-perspective condition ($d = 0.15$, $p = .01$). The degree of bracketing was again higher in the disagreeing-perspective condition: 19% of estimates

bracketed the questions' true answers, compared with both the self-perspective ($d = 0.34$, $p < .001$) and agreeing-perspective ($d = 0.48$, $p < .001$) conditions.

Understanding averaging and bracketing—when is it beneficial? Prior research suggests that bracketing is a key component in understanding why averaging estimates renders an improvement (Larrick & Soll, 2006; Soll & Larrick, 2009). However, as demonstrated by our results on the end-scale questions, this may not always be the case. Specifically, although we observed higher rates of bracketing in the disagreeing-perspective condition for end-scale questions, averaging nonetheless led to a greater overall disadvantage in this condition. To better understand why this occurred, we took a closer look at the underlying components that determine whether averaging first and second estimates is beneficial or not. We formalize each component in Equation 1.

$$
\frac{1}{n}\sum_{i\in n}\left(E_i^{X^f} - E_i^{X^a}\right) = \frac{1}{n}\left[\sum_{i\in n^1}\left(E_i^{X^f} - E_i^{X^a}\right)\right.
$$
$$
+ \sum_{i\in n^2}\left(E_i^{X^f} - E_i^{X^a}\right) + \sum_{i\in n^3}\left(E_i^{X^f} - E_i^{X^a}\right)
$$
$$
\left. + \sum_{i\in n^4}\left(E_i^{X^f} - E_i^{X^a}\right)\right], \tag{1}
$$

where i is the index for individuals; $E_i^{X^f}$ represents error of the first estimate (X^f) of an ith individual on a particular question; $E_i^{X^a}$ represents error of the average estimate (X^a) of an ith individual on a particular question; n is the total set of observations (i.e., number of individuals times the number of questions); n^1 is the subset of observations in n where the second estimate (X^s) moves toward the true answer (X) while $X^f > X^s > X^f − 4\,(X^f − X)$ or $X^f < X^s < X^f + 4\,(X − X^f)$; n^2 is the subset of observations in n in which X^s moves away

Table 4. The Benefit or Disbenefit of Averaging Per Individual Component for All Three Experimental Conditions, Experiment 4 (End-Scale Questions)

Condition	$\frac{1}{n^c} \sum_{i \in n^c} \left(E_i^{X^f} - E_i^{X^a} \right)$	$\frac{1}{n^c} \sum_{i \in n^1} \left(E_i^{X^f} - E_i^{X^a} \right)$	$\frac{1}{n^c} \sum_{i \in n^2} \left(E_i^{X^f} - E_i^{X^a} \right)$	$\frac{1}{n^c} \sum_{i \in n^3} \left(E_i^{X^f} - E_i^{X^a} \right)$	$\frac{1}{n^c} \sum_{i \in n^4} \left(E_i^{X^f} - E_i^{X^a} \right)$
Disagree	−132.20	143.61	−255.68	−17.74	−2.39
Self	2.48	111.52	−105.66	−3.02	−0.36
Agree	−6.71	91.11	−95.03	−2.72	−0.07

Note: Results are shown separately for the disagreeing-perspective (disagree), self-perspective (self), and agreeing-perspective (agree) conditions. See the text for an explanation of the equations.

from X (i.e., $X < X^f < X^s$ or $X > X^f > X^s$); n^3 is the subset of observations in n in which X^s moves toward X while $X^f > X^s < X^f - 4$ $(X^f - X)$ or $X^f < X^s > X^f + 4$ $(X - X^f)$; and n^4 is the subset of observations in n in which $X^f = X$ while $X^s \neq X^f$.

To clarify, the left-hand side of the equation represents the benefit of averaging, and the right-hand side represents the unique components that make up this benefit of averaging. The first component (n^1) represents those observations in the total set of observations (n), where the error of the average estimate is always lower than the error of the first estimate. These observations bring the benefit of averaging estimates (each observation in this subset yields, by definition, a positive number). Here, the second estimate lies in what has been called the "gain range" (Herzog & Hertwig, 2009, p. 232). The other three components (n^2, n^3, and n^4) are those observations where the error of the average estimate is always higher than the error of the first estimate.[6] These observations bring the disbenefit of averaging (each observation in these subsets yields, by definition, a negative number). Averaging first and second estimates (following Equation 1) results in an overall improvement when the part that brings benefit (i.e., the first component) outweighs the parts that bring disbenefit (i.e., the other three components). Likewise, when the parts that bring disbenefit outweigh the part that brings benefit, averaging estimates becomes unbeneficial overall. When we look at each component for the end-scale questions separately per condition (see Table 4; n^c refers to the total set of observations for a particular condition), it becomes clear that the parts that bring disbenefit clearly outweigh the part that brings benefit for the disagreeing-perspective condition (rendering an overall disadvantage of −132.20 in this instance).

What about the observed higher bracketing rate in the disagreeing-perspective condition for end-scale questions? There are two types of brackets (following Equation 1). There are brackets—which we refer to as *beneficial brackets*—in which the average estimate is by definition more accurate than the first estimate (e.g.,

$X = 30$, $X^f = 20$, $X^s = 50$, $X^a = 35$). Beneficial brackets are observations that follow from n^1. *Unbeneficial brackets*, on the other hand, are those observations in which the average estimate is by definition less accurate than the first estimate. Unbeneficial brackets are observations that follow from n^3. These brackets are unbeneficial because the two estimates overbracket a question's true answer, rendering an average estimate that is worse than the first estimate (e.g., $X = 30$, $X^f = 20$, $X^s = 80$, $X^a = 50$). Although these types of brackets are relatively rare, they occurred more frequently in the disagreeing-perspective condition for the end-scale questions (percentage of observations in a condition: disagreeing perspective = 7%; self-perspective = 4%; agreeing perspective = 3%). In sum, although bracketing is indeed a key component when it comes to averaging estimates, it does not by definition render an improvement. That is, averaging estimates becomes unbeneficial once the part that brings benefit (i.e., n^1-observations, including beneficial brackets) is canceled by observations in which the average estimate performs worse than the first estimate (i.e., the n^2, n^3, and n^4 observations).

General Discussion

Many decisions depend on people's ability to make accurate estimates of unknown quantities, and a demonstrated way to improve the accuracy of estimates is to aggregate multiple estimates made by the same person. The potential contained in such an intervention is enormous, and a key challenge is to identify strategies that can help improve the accuracy of people's aggregate estimates (Herzog & Hertwig, 2014a). In this article, we introduced the following strategy: Combine people's first estimate with their second estimate made from the perspective of someone they often disagree with. Across five experiments, we found evidence that such a strategy produces accurate estimates. These results underscore the importance of perspective taking and disagreement as strategies to improve the accuracy of people's quantitative estimates.

The presented findings indicate the benefits of disagreement, a component of people's social interactions that is usually presented as undesirable (Kennedy & Pronin, 2008; Reeder et al., 2005; Sunstein, 2002). What is particularly interesting is that people obtained more accurate estimates by changing their perspective. It remains to be seen whether taking the perspective of any other people—say, experts in a particular field—would lead to similar benefits. This might be an important future research direction, as our findings demonstrate that taking the perspective of other people (e.g., an agreeing perspective) might not always render an increase in accuracy compared with simply making a second guess.

Although the inner crowd offered a gain in accuracy, we also identified a situation in which it backfired, leading to no improvement or even worse performance. We found this to be the case when a question's answer was close to the scale's end. Importantly, for participants who employed the disagreeing-perspective strategy, the accuracy of their average estimate was much worse than their first estimate for these types of questions. What is particularly interesting is that the propensity of people to move away from the answer when making second estimates is introduced through a feature of the situation rather than some innate bias (Gaertig & Simmons, 2021; Herzog et al., 2019; Müller-Trede, 2011).

Our research also has several limitations. First, the presented evidence is restricted to populations from the United States and United Kingdom, and future work needs to confirm whether these findings hold true in other parts of the world. Second, although combining initial estimates with second estimates made from a disagreeing perspective is beneficial, the presented research remains mute as to whether people would be willing to aggregate both estimates when given the opportunity (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014b; Müller-Trede, 2011). Although prior work indicates that people are more likely to combine their estimates when they actively opposed themselves through dialectical bootstrapping (Herzog & Hertwig, 2014b), it remains to be seen whether this holds true when the opposition comes from someone with whom they often disagree. People typically view others holding opposing views and opinions less favorably (Iyengar & Westwood, 2015; Kennedy & Pronin, 2008; Reeder et al., 2005), potentially undermining their willingness to include the viewpoints of disagreeing others into their own judgments. Future research could address this issue in more detail by testing under what conditions people are willing to combine their estimates with the estimates of disagreeing others to obtain more accurate estimates.

On a final note, whereas previous studies often relied on natural processes such as forgetting or the passage of time to improve the accuracy of inner crowds, the present findings report a strategy that is more convenient and time efficient. Similar to other, more active interventions (Herzog & Hertwig, 2009; Litvinova et al., 2020; Winkler & Clemen, 2004), taking a disagreeing perspective can likewise be used as a potent strategy when people cannot benefit from the wisdom of an actual crowd. Overall, combining one's first estimate with a second estimate made from the perspective of disagreeing others proves to be a convenient and effective judgment tool.

ORCID iDs

Philippe P. F. M. Van de Calseyde ![ORCID] https://orcid.org/0000-0003-0566-7018
Emir Efendić ![ORCID] https://orcid.org/0000-0002-2365-0247

Notes

1. Here, we kept the sign of the error because the size as well as the direction of the error are informative.
2. Following prior work on the inner crowd, we initially preregistered an intention to conduct simple *t* tests rather than mixed-effect models for Experiments 1a, 2, and 4. We made this

change in response to a suggestion during the review process.
3. Using the mean absolute error produces the same results qualitatively. Mean-absolute-error results for all experiments can be found at https://osf.io/ewpyq/.

4. Note that in Experiment 1b, we also measured the time participants needed to generate their second estimates. Comparing this time between the self- and disagreeing-perspective conditions showed that there was no difference, $d = 0.08$, $p = .21$, Bayes factor favoring the null over the alternative hypothesis $(\mathrm{BF}_{01}) = 8.78$.

5. When multiple experiments are conducted, the presence of some nonsignificant findings is to be expected given the nature of hypothesis testing (Lakens & Etz, 2017). To assess the overall evidential value of the prediction that the benefit of averaging is higher when one takes a disagreeing perspective, we aggregated the data of the same six questions from Experiments 2, 3, and 4. Results showed that, overall, participants in the disagreeing-perspective condition benefited more from averaging than participants in the agreeing-perspective condition $(d = 0.18, p < .001)$.

6. Note that there are also observations where the error of the average is identical to the error of the first estimate—that is, observations where X^{s} moves toward X while $X^{\mathrm{f}} > X^{\mathrm{s}} = X^{\mathrm{f}} - 4 (X^{\mathrm{f}} - X)$ or $X^{\mathrm{f}} < X^{\mathrm{s}} = X^{\mathrm{f}} + 4 (X - X^{\mathrm{f}})$, and observations where $X^{\mathrm{s}} = X^{\mathrm{f}} = X^{\mathrm{a}}$. These observations are not included in the equation because including them does not render any benefit or disbenefit.

References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., & Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PLOS ONE*, *14*(12), Article e0225826. https://doi.org/10.1371/journal.pone.0225826

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130–147.

Baars, J. A., & Mass, C. F. (2005). Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics. *Weather and Forecasting*, *20*(6), 1034–1047.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin*, *23*(7), 751–758.

Central Intelligence Agency. (2020). *The World Factbook*. https://www.cia.gov/the-world-factbook/

Central Intelligence Agency. (2021). *The World Factbook*. https://www.cia.gov/the-world-factbook/

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

de Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences, USA*, *115*(9), 2066–2071.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17–38.

Gaertig, C., & Simmons, J. P. (2021). The psychology of second guesses: Implications for the wisdom of the inner crowd. *Management Science*, *67*(9), 5301–5967. https://doi.org/10.1287/mnsc.2020.3781

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708–724.

Galton, F. (1907). The ballot-box. *Nature*, *75*(1952), 509–510.

Gordon, M., Viganola, D., Dreber, A., Johannesson, M., & Pfeiffer, T. (2021). Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE*, *16*(4), Article e0248780. https://doi.org/10.1371/journal.pone.0248780

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. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506.

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 218–232.

Herzog, S. M., Litvinova, A., Yahosseini, K. S., Tump, A. N., & Kurvers, R. H. J. M. (2019). The ecological rationality of the wisdom of crowds. In R. Hertwig, T. J. Pleskac, T. Pachur, & The Center for Adaptive Rationality (Eds.), *Taming uncertainty* (pp. 245–262). MIT Press.

Hoever, I. J., Van Knippenberg, D., Van Ginkel, W., & Barkema, H. G. (2012). Fostering team creativity: Perspective taking as key to unlocking diversity's potential. *Journal of Applied Psychology*, *97*(5), 982–996.

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

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities

for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068–1074. https://doi.org/10.1037/a0019694

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*, 690–707.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, *107*(2), 384–396.

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. https://doi.org/10.1177/0956797620941840

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, *34*(6), 833–848.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lakens, D., & Etz, A. J. (2017). Too true to be bad. *Social Psychological and Personality Science*, *8*(8), 875–881.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127.

Lewis, J., Gaertig, C., & Simmons, J. P. (2019). Extremeness aversion is a cause of anchoring. *Psychological Science*, *30*(2), 159–173. https://doi.org/10.1177/0956797618799305

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (2020). How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision*, *7*(3), 183–211.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, *6*(4), 283–294.

Mutz, D. C. (2006). *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

Piaget, J. (1965). *The moral judgement of the child*. Free Press. (Original work published 1932)

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin*, *31*(11), 1498–1510.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805.

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*, *5*, 786–794. https://doi.org/10.3389/fpsyg.2014.00786

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, *15*(5), 550–562. https://doi.org/10.1037/h0070482

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, *10*(2), 175–195.

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

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Van Dolder, D., & Van den Assem, M. J. (2018). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behaviour*, *2*(1), 21–26.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, *1*(3), 167–176. https://doi.org/10.1287/deca.1030.0008