http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–22
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Psychology of Second Guesses: Implications for the Wisdom of the Inner Crowd

Celia Gaertig,[a] Joseph P. Simmons[b]

[a] Booth School of Business, University of Chicago, Chicago, Illinois 60637; [b] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19148

Contact: celia.gaertig@chicagobooth.edu, https://orcid.org/0000-0002-5444-4999 (CG); jsimmo@upenn.edu (JPS)

Received: May 6, 2018
Revised: May 14, 2019; May 5, 2020
Accepted: July 14, 2020
Published Online in Articles in Advance: March 23, 2021

https://doi.org/10.1287/mnsc.2020.3781

Copyright: © 2021 INFORMS

Abstract. Prior research suggests that averaging two guesses from the same person can improve quantitative judgments, a phenomenon known as the "wisdom of the inner crowd." In this article, we find that this effect hinges on whether people explicitly decide in which direction their first guess had erred before making their second guess. In nine studies ($N = 8{,}465$), we found that asking people to explicitly indicate whether their first guess was too high or too low before making their second guess made people more likely to provide a second guess that was more extreme (in the same direction) than their first guess. As a consequence, the introduction of that "Too High/Too Low" question reduced (and sometimes eliminated or reversed) the wisdom-of-the-inner-crowd effect for (the majority of) questions with non-extreme correct answers and increased the wisdom-of-the-inner-crowd effect for questions with extreme correct answers. Our findings suggest that the wisdom-of-the-inner-crowd effect is not inevitable but rather that it depends on the processes people use to generate their second guesses.

History: Accepted by Yuval Rottenstreich, decision analysis.
Funding: Financial support was provided by the Beatrice Foods Co. Faculty Research Fund at the University of Chicago Booth School of Business and the Wharton Behavioral Laboratory at the University of Pennsylvania.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/mnsc.2020.3781.

Keywords: wisdom of crowds • estimation • crowd within • debiasing • intuitive confidence

Introduction

Quantitative judgments pervade the business world. Important managerial and consumer decisions hinge on accurate forecasts of political events, market conditions, the likelihood of a product's success, and so on. It is not surprising, then, that researchers have expended considerable effort trying to identify practical ways to improve quantitative judgments.

One of the easiest ways to improve quantitative judgments involves averaging the guesses of multiple judges. Decades of research demonstrate, across a wide variety of tasks and domains, that these averages typically outperform the single guess of one person (Galton 1907, Treynor 1987, Ariely et al. 2000, Surowiecki 2004, Yaniv 2004, Mannes et al. 2014), even when that person has a very good track record (Mannes et al. 2014). This *wisdom-of-crowds* effect emerges because averaging helps to cancel out idiosyncratic biases and random errors.

Recently, researchers have shown that a similar, albeit smaller benefit can be accrued by averaging two guesses from the *same* individual, an effect dubbed the "wisdom of the *inner* crowd" (Ariely et al. 2000; Winkler and Clemen 2004; Vul and Pashler 2008; Herzog and Hertwig 2009, 2014a, b; Steegen et al. 2014; van Dolder and van den Assem 2018). This work suggests that the average of a person's first and second guesses is usually at least slightly better than the person's first guess. The implications of this finding are important, as it suggests that one can make better estimates and predictions simply by making two sequential judgments and averaging them. It is hard to imagine an easier-to-implement prescription for improving one's forecasts.

In this article, we investigate how the wisdom-of-the-inner-crowd effect is affected by whether, in the process of making a second guess, people consciously consider the direction in which their first guess was wrong.

How Do People Generate Second Guesses?

To understand what causes the wisdom-of-the-inner-crowd effect, it is first important to consider the math of it. For the average of two guesses to outperform one's first guess, two conditions must hold, neither of which is inevitable: (1) one's second guess must be in the right direction with respect to her first guess, and (2) one's second guess must not be too far in the right direction. For example, if someone gives a first guess of 45 to a question with a correct answer of 50, the wisdom-of-the-inner-crowd effect can emerge

only if the second guess is (1) greater than 45 (i.e., in the right direction) and (2) less than 65 (i.e., not too far in that direction).

It is also important to consider how people might generate their second guesses, for how they are generated may influence the wisdom of the inner crowd. Here we will consider two different possibilities, a resampling process, in which people do not explicitly decide in which direction their first guess had erred, and a choice process, in which people do. As we will see, the wisdom of the inner crowd may hinge on which of these processes people actually use to generate their second guesses.

Resampling Process

According to past research, the wisdom-of-the-inner-crowd effect emerges because averaging cancels out the random errors that pervade people's judgments. By this account, people are presumed to sample their first guess from an internal distribution of possible guesses and to then resample their second guess from that same distribution (Vul and Pashler 2008, Hourihan and Benjamin 2010).

To understand how such a resampling process will generate the wisdom of the inner crowd, consider an example. Imagine that a participant named Sally is tasked with estimating the percentage of a survey's responders who prefer psychology to economics and that the true answer is 55%. For simplicity, assume that Sally correctly believes, *on average*, that the true answer is 55%, but that she is, like all humans, inconsistent, meaning that her guesses are subject to the vagaries of random error (Dawes 1979, Ashton 2000). Thus, on one occasion she might guess 55%, on another she might guess 45%, on another she might guess 65%, and so on. To make this concrete, say that Sally's first guess is randomly drawn from a normal distribution with a mean of 55% and a standard deviation of 10% and that her first guess happens to be 62% (Figure 1(a)).

Now, if Sally engages in a resampling process, the process she uses to generate her second guess will be exactly the same as the process she used to generate her first guess: She will again randomly sample from a normal distribution with a mean of 55% and a standard deviation of 10%. In Figure 1(a), it is not hard to see that, in this case, Sally's second guess is much more likely than not to be in the right direction with respect to her first guess and that her average guess is much more likely than not to be better than her first guess. The wisdom-of-the-inner-crowd effect is likely to emerge.[1]

Thus, people may exhibit a tendency to adjust their first guesses in the right direction without explicitly considering the direction in which their first guesses had erred. In other words, they may unconsciously adjust in the right direction more often than not, simply because their first and second guesses are randomly drawn from similar distributions. If people are using this kind of resampling process to generate their second guesses, then, over a large sample of judgments, the wisdom-of-the-inner-crowd effect is virtually inevitable.

Choice Process

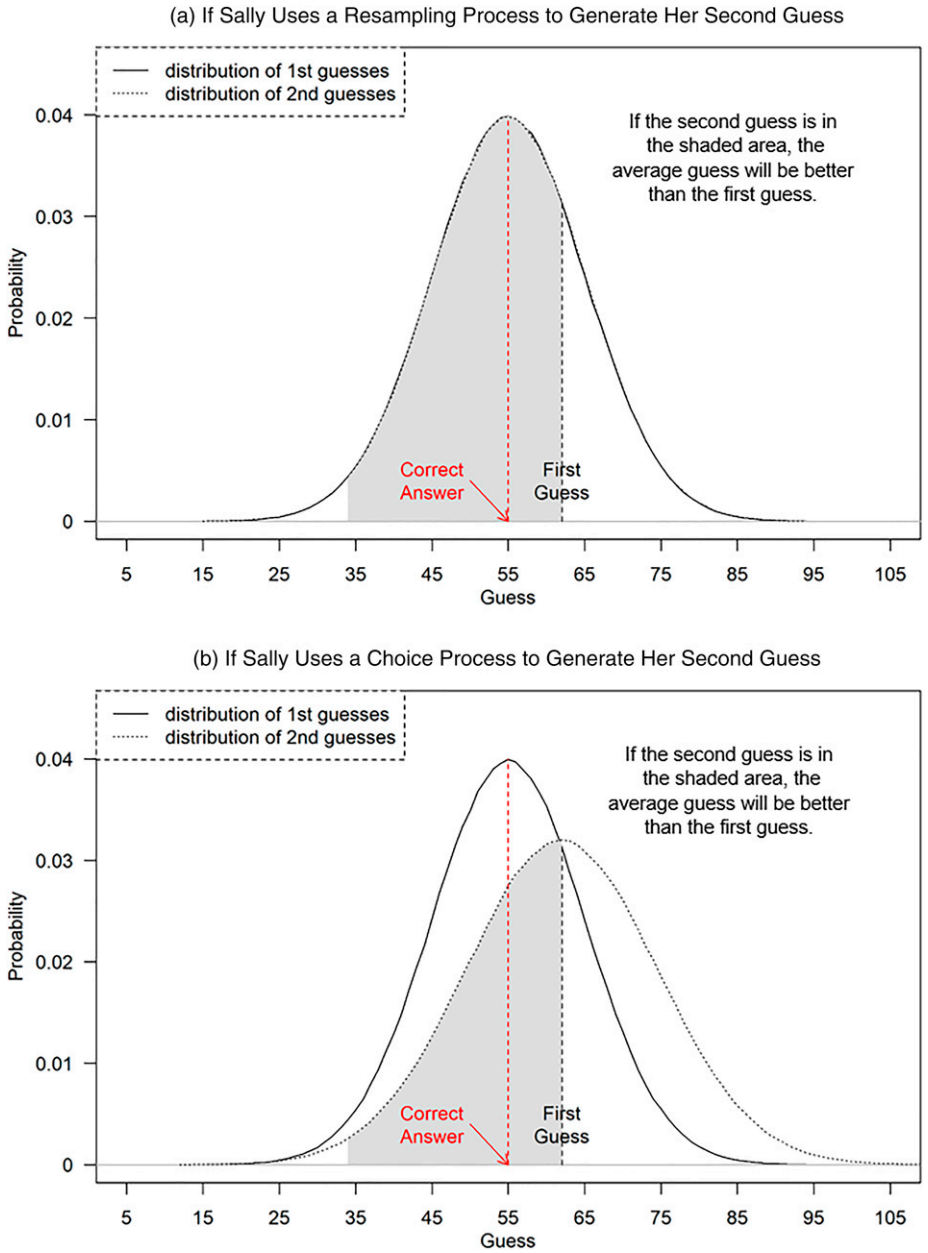
However, this resampling process may not fully capture how people actually generate their second guesses. It is not farfetched to imagine that some people, in some circumstances, may instead adopt a more explicit choice process, one that involves consciously deciding in which direction to adjust their first guess. How *this* process affects the wisdom of the inner crowd depends on exactly how someone makes this choice. However, for now, let us return to the example of Sally and focus on the plausible possibility that explicitly considering the direction in which her first guess had erred makes her realize that she *does not know* in which direction her first guess had erred. Indeed, her first guess may have been her first guess precisely because it seemed like the best possible answer. In this case, she may randomly decide on a direction of adjustment, effectively drawing her second guess from a new distribution that is *centered on her first guess*.

Figure 1(b) shows what happens if *this* is how Sally generates her second guess. As you can see, if people's second guesses are randomly drawn from a distribution centered on their first guess, then they will adjust in the correct direction only 50% of the time, and thus we would never expect the average of a person's first and second guesses to have more than a 50% chance to outperform her first guess. Rather, because at least some percentage of guesses that are in the right direction would be *too far* in that direction, we would expect to observe the opposite: an *ignorance*-of-the-inner-crowd effect (see the simulations in the e-companion).

To get an intuition for why this is, consider again that the wisdom-of-the-inner-crowd effect can only emerge if (1) Sally adjusts in the right direction relative to her first guess, and (2) Sally does not adjust too far in the right direction. If Sally *randomly* chooses in which direction to adjust her second guess, then she will be adjusting in the right direction only 50% of the time, and so her average cannot outperform her first guess more than 50% of the time. Moreover, because at least some of her right-direction second guesses will be too far in the right direction, we should expect most of her first guesses to outperform her average guesses.

From this exercise, we see that it is at least theoretically possible for a reliance on a choice process to eliminate or even reverse the wisdom-of-the-inner-crowd effect. However, whether it actually does so depends on exactly how that choice process unfolds. In our illustration,

Figure 1. Example of How Using a Resampling Process (a) or Choice Process (b) to Generate Second Guesses Can Affect the Wisdom of the Inner Crowd



Notes. The probability of making a second guess that produces a wisdom-of-the-inner-crowd effect is 0.64 in (a) and 0.44 in (b). The R code to reproduce this figure is here: https://researchbox.org/45.

we imagined that a person decided whether her first guess was too high or too low by choosing randomly. However, there are of course other possibilities. For example, perhaps explicitly considering the direction in which they had previously erred makes people reflect and subsequently identify the nature of their mistake, making them even more likely to provide a second guess that is in the right direction. In that case, the adoption of a choice process might improve rather than worsen the wisdom of the inner crowd.

Indeed, the key takeaway from this section is not that a choice process will inevitably decrease the wisdom-of-the-inner-crowd effect but rather that it *may* affect it. Whether, and how, it affects it is ultimately an empirical question, one that we attempt to answer in the remainder of this article.

The Current Research

In what follows, we report the results of experiments that examine what happens to the wisdom-of-the-inner-crowd effect when, just before making their

second guess, people consider the direction in which their first guess had erred. Specifically, we investigated whether asking participants to explicitly indicate whether their first guess was too high or too low affects the wisdom of the inner crowd. For example, in Study 1, we asked participants to estimate what percentage of survey responders would express a particular preference. Shortly after participants made their first guess, we reminded them of that guess and asked them to make a second, different guess. As illustrated in Figure 2, this is the point at which we introduced our key experimental manipulation. Before making their second guess, participants randomly assigned to the "Too High/Too Low" condition were asked to explicitly decide whether their first guess was too high or too low (hereafter referred to as the "Too High/Too Low question"), and participants randomly assigned to the "Estimate Only" condition were not. There were no other differences between the conditions.

This design allowed us to test whether a reliance on a choice process (in the Too High/Too Low condition) would influence the wisdom of the inner crowd. Before investigating this question, we had no firm predictions about whether, or how, this would affect the wisdom of the inner crowd.

First, it is reasonable to imagine that asking participants to decide whether their first guess was too high or too low would *benefit* the wisdom of the inner crowd. In fact, prompting participants to think about whether their first guess was too high or too low is part of the *dialectical bootstrapping* instructions (Herzog and Hertwig 2009, 2014a, b) that have been proposed as a way to increase the wisdom of the inner crowd. Specifically, Herzog and Hertwig (2009, 2014a, b) found that the wisdom of the inner crowd can be enhanced by asking participants to think about why their first guess might have been wrong using the following four steps in their thought process:[2]

1. For a moment, assume that your estimate is "wrong" or clearly off the mark.

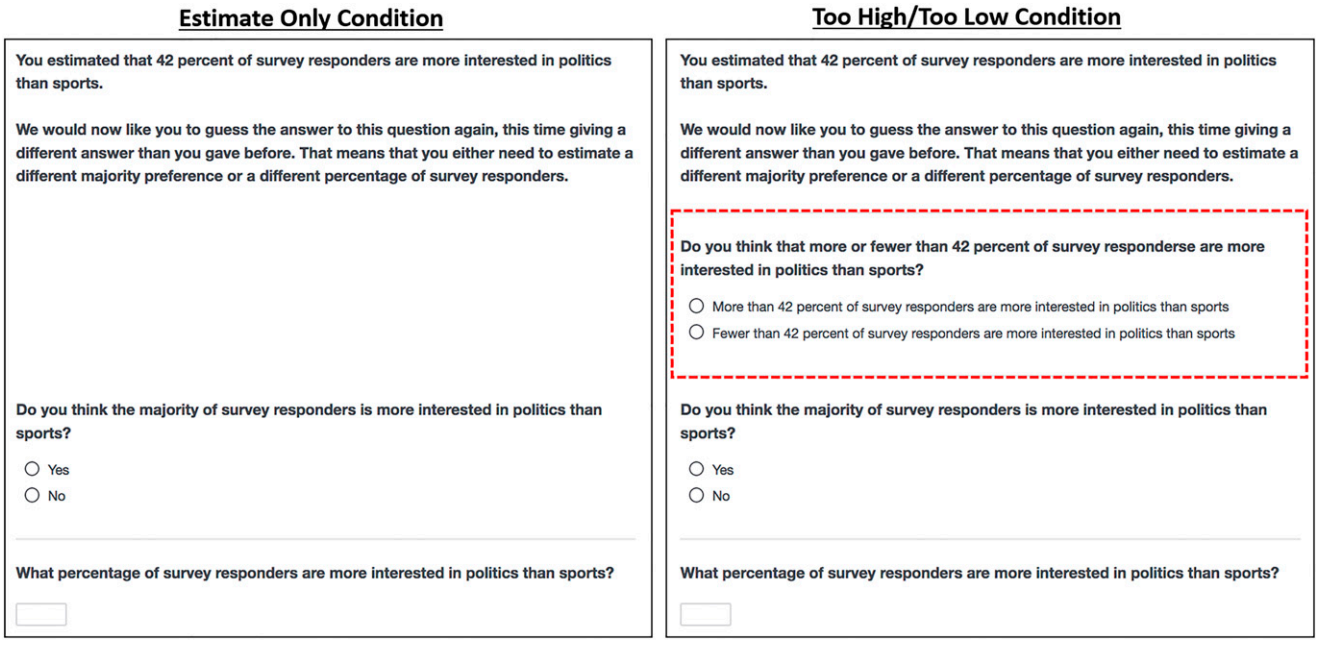
2. Think about a few reasons why your original estimate could be clearly off the mark. Which assumptions and considerations that led to your original estimate could be wrong?

3. What do these new considerations and the resulting different and/or new assumptions imply? Was your original estimate rather too high or too low? Please decide on "rather too high" or "rather too low".

4. Based on this "new" alternative view (with different and/or new assumptions, knowledge, etc.), please make a new, second, alternative estimate.

As can be seen from these instructions, the third step asks participants to think about whether their first estimate was too high or too low and to decide on the direction. Although participants in experiments of Herzog and Hertwig (2009, 2014a, b) were not asked to explicitly answer this question, their results suggest that prompting people to consider the direction in which their first guess had erred might increase their likelihood of giving second guesses that are in the correct direction and thereby improve the wisdom of the inner crowd.

Figure 2. An Example of the Manipulation Used in Our Studies



Note. The only difference between the two experimental conditions was whether participants were asked to answer the Too High/Too Low question (framed with a dashed red line) before making their second guess.

Alternatively, and as described previously, it is also reasonable to imagine that asking people to decide whether their first guess was too high or too low would hurt the wisdom of the inner crowd. This would be the case if answering the Too High/Too Low question disrupts the resampling process and induces the kind of random-direction choice process that we spelled out in the previous section.

We conducted nine experiments in which we asked people to make two different (and usually incentivized) guesses of some unknown quantity. In Studies 1–4, 6, 7, and S1, participants were asked to estimate the preferences and behaviors of other survey responders; in Study 5, participants were asked to predict the outcomes of upcoming NBA basketball games; and, in Study 8, participants were asked to predict stock prices. In our studies, we examined whether asking participants to explicitly answer the Too High/Too Low question before making their second guess affects the probability (1) that the average guess outperforms the first guess and (2) that the second guess is in the right direction relative to the first guess. After observing a consistent pattern of results in Studies 1–5, we conducted Study 6 to explore the mechanism of our effects and Studies 7, 8, and S1 to explore the specificity and robustness of our effects.

We report all our measures, manipulations, exclusions, and how we determined our sample sizes. We preregistered all our studies, and we provide the link to the preregistrations in the Appendix. All our data and materials are available at this link: https://researchbox.org/45. Additional analyses are provided in the e-companion (see Table A1 for the table of contents of the e-companion).

Studies 1–5

In Studies 1–4, we asked participants to predict the percentage of survey responders who would report having a certain preference or engaging in a particular behavior. In Study 5, we asked participants to predict the point differentials of upcoming NBA games. Because the methods and results of these five studies are similar, we describe them all at once.

Method

Participants. We conducted Studies 1 and 3–5 on Amazon's Mechanical Turk (MTurk). Participants received $1 for participation. In Studies 1 and 3, participants could also earn up to an additional $1 for accurate prediction performance, and in Study 5, participants could earn up to an additional $0.80 for accurate prediction performance. For Studies 1, 3, 4, and 5, we preregistered to collect data from 600, 1,300, 600, and 600 participants, respectively, and we wound up with final samples of 604, 1,260, 585, and 610 participants, respectively. The final sample for each study included all participants who made a first and second guess for at least one of the prediction questions while excluding those whom we preregistered to exclude.[3] The samples averaged 33–37 years of age. The samples of Studies 1, 3, and 4 were 53%–59% female, and Study 5's sample (the NBA basketball study) was 37% female.

Study 2 was conducted in the laboratory. Participants received a $10 show-up fee for a one-hour-long laboratory session of which ours was a 10-minute component. We preregistered to run as many laboratory sessions as needed to get at least 400 participants, and we ended up with a final sample size of 457 participants (average age = 21 years; 58% female). Our final sample included all participants who made a first and second guess for at least one of the prediction questions and excluded duplicate responses from six participants whose laboratory ID appeared twice in the data set.[4]

Procedures. ***Overview.*** This section describes the procedures of each of the five studies, beginning with a detailed description of Study 1 and then briefer descriptions of the ways in which Studies 2–5 differed from Study 1.

Study 1. In Study 1, participants estimated the preferences and behaviors of all people completing the study. At the beginning of the survey, participants answered 10 questions about their own preferences and behaviors, each of which involved a binary choice (e.g., "What are you more interested in: politics or sports?"). Asking participants to provide their own preferences and behaviors was necessary in order to determine the true answers for this study. The 10 questions were presented to participants in a random order on the same screen.

Next, participants were asked to estimate what percentage of survey responders would give a particular answer to the same 10 questions. Table 1 displays the 10 prediction questions and their true percentages in Studies 1–4. For each of the 10 preference/behavior questions participants were asked to estimate (1) whether the majority of survey responders would report having the preference or engaging in the behavior (e.g., "Do you think the majority of survey responders is more interested in politics than sports?"), and (2) what the percentage was (e.g., "What percentage of survey responders are more interested in politics than sports?"). The wording of question (2) was the same for all participants, regardless of how participants answered question (1). For example, all participants were asked, "What percentage of survey responders are more interested in politics than sports?," regardless of whether they previously indicated that the majority is more interested in politics or that the majority is more interested

Table 1. Prediction Questions and True Answers in Studies 1–4

Percentage of participants who	Studies 1 and 2	Study 3	Study 4
…prefer Indian food to Mexican food	16%	17%	18%
…prefer having a cat to having a dog	34%	36%	36%
…own an iPad	43%	41%	41%
…prefer vanilla ice cream to chocolate ice cream	51%	48%	50%
…have posted a video on YouTube	51%	54%	59%
…drink at least 10 cups of coffee in a week	54%	45%	44%
…prefer Spring to Summer	63%	64%	64%
…prefer politics to sports	64%	59%	59%
…have a Twitter account	65%	70%	69%
…have traveled outside the United States	69%	71%	74%

Notes. The true answers for Studies 1 and 2 are identical, because Study 2 was conducted in the laboratory, and we asked participants to predict the preferences and behaviors of survey responders who previously completed the study online (i.e., those of Study 1 participants).

in sports. Participants made their guesses for each of the 10 prediction questions on the screen one at a time, and the order in which the prediction questions were presented was randomized between subjects.

After making their first guesses for all 10 prediction questions, participants were (unexpectedly) asked to make a second, different guess for each of the questions. We told participants, "On the next pages, we will ask you to estimate the behaviors and preferences of people completing this survey again. We would like you to make a second estimate that is not exactly the same as your first estimate." Participants were then presented with each of the prediction questions again (in a random order), with each question on its own page. For each question, they were shown their first guess and were reminded to provide a different second guess. They were then asked to again indicate (1) whether the majority of survey responders would report having the preference or engaging in the behavior and (2) what the percentage was.

As described previously (and illustrated in Figure 2), we manipulated (between subjects) whether we asked participants to decide whether their first guess was too high or too low right before they made their second guess. Participants in the Too High/Too Low condition were asked to decide whether their first guess was too high or too low (e.g., "Do you think that more or fewer than [first guess] percent of survey responders are more interested in politics than sports?"). Participants in the Estimate Only condition were not asked the Too High/Too Low question but were only asked to make a second guess. As Figure 2 shows, whether the Too High/Too Low question was asked was the only difference between the two conditions.[5]

In Studies 1 and 5, we included an additional condition in which participants were asked the Too High/Too Low question but were not asked to make a second guess immediately after seeing the Too High/Too Low question. Participants in this condition made their second guess in a separate block at the end of the study, after they answered the Too High/Too Low question for all prediction questions. We present the results from this condition in the e-companion.[6]

Incentives. Participants were incentivized to make accurate predictions. They received \$0.10 for each percentage estimate that was within 1 percentage point of the true answer (i.e., up to \$1 in total). We used the better of participants' two guesses for each prediction question to determine whether they won a bonus. Before participants started making their second guesses, we told them that making a second guess meant having a second chance to win a bonus.

Study 2. Study 2 examined whether the results from Study 1 would replicate in the laboratory. Study 2 followed the same procedure as Study 1 except for three changes. First, participants predicted the preferences/behaviors of people who previously completed the study online (i.e., those of Study 1 participants). This enabled us to calculate bonus payments for each participant while they were taking the study so that we could pay them at the end of the laboratory session (rather than needing to wait until data collection for Study 2 was completed). Second, we changed the bonus payment rule so that we provided a bonus of \$1 to participants if, across all 10 items, the better of their two guesses was on average less than 10 percentage points away from the true answers. Thus, participants either received no bonus or a \$1 bonus. Third, in addition to telling participants that making a second guess meant having a second chance to win a bonus, we also explicitly told participants that the better of their two guesses would determine their bonus payment.

Study 3. Study 3 examined whether the results from Study 1 would be robust to changes to the incentive structure. Study 3 followed the same procedure as Study 1 except that we manipulated whether

participants' bonuses were based on the better of participants' two guesses or on only their second guesses. Before making their first guesses, all participants were told, "Please try your best to be accurate." We introduced our incentive manipulation just before participants made their second guesses. Participants in the Better Guess Bonus condition were told that we would use the best of their two estimates to determine their bonus payment. Participants in the Second Guess Bonus condition were instead told that we would only use the estimate that they are about to make (i.e., their second estimate) to determine their bonus payment. We then asked participants a comprehension check question to ensure that they understood which of their guesses would determine their bonus payment. Participants were given two tries to answer this question correctly. Only 38 of the 1,260 participants (3% of participants) failed to answer the comprehension check question correctly after two tries. We preregistered to retain these participants in the analyses, but our findings do not change if we exclude them.[7]

Study 4. Study 4 provided an additional test of the robustness of our results to changes to the incentives structure. Study 4 followed the same procedure as Study 1 except that we did not provide any monetary incentives for accurate prediction performance. Participants were simply told to try their best to be accurate.

Study 5. In Study 5, we examined whether the results of the previous studies would replicate in a different forecasting domain. Participants were asked to predict the outcomes of eight upcoming NBA games. We selected eight games from those that were played on February 23 and 24, 2017. We posted the study one day before the first game day. For each game, participants were presented with the game's start time, the names of the home and visiting teams, and the teams' win-loss records. The order of presentation of the games was randomized between subjects, and each game was presented to participants on the screen one at a time.

For each game, participants were asked to predict which team would win and by how many points.[8] After making their predictions for all eight games, participants were (unexpectedly) asked to predict the outcomes of the games again. We told participants, "We would now like you to make a second prediction that is not exactly the same as your first prediction. This means that you must either predict a different winner or a different point differential." Participants were then presented with each of the games again (in a random order), with each game on its own page. For each game, they were shown their first guess and were reminded to provide a second, different guess by indicating again who would win the game and by how many points.

As in Studies 1–4, we manipulated between subjects whether we asked participants to decide whether they thought that their first guess was too high or too low before they made a second guess. Specifically, participants in the Too High/Too Low condition were asked to indicate whether they thought that their predicted winning team would defeat their predicted losing team by more than they had predicted or not. For example, if a participant's first guess was that the Mavericks would defeat the Timberwolves by 2 points, they were asked, "If you were to guess again, do you think that the Dallas Mavericks will defeat the Minnesota Timberwolves by more than 2 point(s) or not?" Participants were then asked to make their second guess. Participants in the Estimate Only condition were not asked the Too High/Too Low question but were only asked to make a second guess.

Incentives. As in Study 1, participants were incentivized to make accurate predictions, and the better of participants' two guesses was used to determine whether they won a bonus. Participants received \$0.10 for each of the eight game predictions that they got exactly right; because they predicted the outcomes of eight games, they could earn up to an additional \$0.80.

Primary Dependent Measures. We focus our analyses of Studies 1–5 on three dependent measures. In this section, we describe how we operationalized these measures for each prediction domain.

The first two measures were preregistered in all of our studies and assessed whether the average of participants' first and second guesses outperformed their first guess and whether participants' second guess was in the right direction relative to their first guess. The third measure was preregistered in some of our studies and gave us insight into why our effects emerge.

1. Did participants' average guesses outperform their first guesses? Our first dependent measure was a binary variable indicating whether, for each observation, the participant's average guess outperformed her first guess.[9] In Studies 1–4, we defined accuracy relative to the true percentage. In Study 5, we defined accuracy relative to the game's point spread, because the game's point spread is the (wise) prediction made by well-calibrated betting markets and is less susceptible to chance factors than are actual game outcomes (Kelly and Simmons 2016).

The average guess outperformed the first guess if it was closer to the true percentage/point spread than the first guess was. For example, in Studies 1–4, if a participant's first guess was off by 22, then the average guess outperformed the first guess if it was off by less than 22. Similarly, in Study 5, if a participant's first guess was that Team A would win by 2, but the point spread predicted that Team B would win by 4,

then the participant's first guess was off by 6 points, and the average guess outperformed the first guess if it was off by less than 6 points.

2. Were participants' second guesses in the right direction relative to their first guesses? Our second dependent measure was a binary variable indicating whether, for each observation, the participant's second guess was in the right direction relative to her first guess. In Studies 1–4, we defined *right direction* with respect to the true percentage, and in Study 5, we defined *right direction* with respect to the game's point spread.

A participant's second guess was in the right direction if it was in the direction of the correct answer with respect to the first guess. For example, in Studies 1–4, if a participant's first guess was 42% and the true answer was 64%, then any second guess greater than 42% was in the right direction. Similarly, in Study 5, if a participant's first guess was that Team A would win by 2 and the point spread predicted that Team B would win by 4, then any second guess that predicted that Team A would win by less than 2 or that Team B would win by any amount was a second guess in the right direction.

3. Were participants' second guesses more extreme in the same direction than their first guesses? This binary variable indicated whether, for each observation, the participant's second guess was more extreme in the same direction than her first guess. We defined *more extreme* relative to a participant's first guess and the midpoint of the scale. In Studies 1–4, the midpoint of the scale was 50%, and in Study 5, we presumed the psychological midpoint of the scale to be a tied game (i.e., a point differential of zero).

A participant's second guess was considered more extreme in the same direction than their first guess if it was further away from the midpoint of the scale in the same direction than their first guess was. For example, in Studies 1–4, if a participant's first guess was 42%, then any second guess below 42% was coded as a more extreme second guess. Similarly, in Study 5, if a participant's first guess was that Team A would win by 2, then any second guess that predicted that Team A would win by more than 2 was coded as a more extreme second guess.

We ran a few studies before discovering that this extremity measure would help us understand why our manipulation affects the wisdom of the inner crowd. We did not preregister this measure in Studies 1, 2, and 5, which were run before Studies 3, 4, 6, 7, and 8. However, we preregistered this measure in our other studies, and we consistently found the same effects across all of them.

Other Preregistered Measures. In some of our studies, we also preregistered to analyze a few other measures. In Studies 2–4, 6, and 7, we preregistered to analyze the absolute difference between first and second guesses and the average within-subject correlation between first and second guesses. We report the results of these measures below (for Studies 1–5). In Study 2, we also preregistered to analyze a binary variable indicating whether participants *switched sides* when making their second guess (i.e., whether their second guess was on a different side of 50% than their first guess was). This switching sides measure was a precursor to our (superior) measure that assessed whether participants made more extreme second guesses, and thus we do not discuss it further.

Results and Discussion

Analysis Plan. In Studies 1–4, each participant who fully completed the study contributed 10 rows to the data set, one for each of the preferences/behaviors they estimated. In Study 5, each participant who fully completed the study contributed eight rows to the data set, one for each of the games they predicted. From the data sets of Studies 1, 3, 4, and 5, we dropped 16, 21, 28, and 23 observations, respectively, because the participant's second guess was missing, preventing us from being able to calculate our dependent measures. From the data sets of Studies 1 and 3, we additionally dropped 11 and 2 observations, respectively, because the first and second guesses were identical.[10]

We analyzed the data from each of the studies separately. In Studies 1, 2, 4, and 5, we regressed the dependent measures on the Too High/Too Low condition,[11] and, in Study 3, we regressed the dependent measures on (1) the Too High/Too Low condition (contrast-coded), (2) the Second Guess Bonus condition (contrast-coded), and (3) their interaction. As preregistered, we present the results from ordinary least squares (OLS) regressions; logistic regressions yielded nearly identical results. All of our regression analyses included fixed effects for item. We clustered standard errors by participant to account for the nonindependence of observations.

Main Analyses. As shown in Table 2, answering the Too High/Too Low question decreased the percentage of observations for which the average guess outperformed the first guess. Indeed, while we tended to observe a small wisdom-of-the-inner-crowd effect in the Estimate Only condition, we tended to observe a small *ignorance*-of-the-inner-crowd effect in the Too High/Too Low condition. This reduction of the wisdom of the inner crowd was directionally observed in all five studies; it was statistically significant in three of them (Studies 1, 2, and 5) and marginally significant in another (Study 3).

Table 2 also shows that answering the Two High/Too Low question decreased the likelihood that participants' second guesses would be in the right direction relative to their first guesses. This effect was statistically

Table 2. Results for Studies 1–5

Prediction domain	Study	Estimate Only	Too High/Too Low	Effect of the Too High/Too Low condition (vs. Estimate Only condition)
% of observations for which the average guess outperformed the first guess				
Preferences	Study 1	51.1%	46.5%	$b = -0.046$, SE = 0.018, $p = 0.011$
	Study 2	55.2%	49.1%	$b = -0.061$, SE = 0.015, $p < 0.001$
	Study 3	51.4%	49.6%	$b = -0.017$, SE = 0.010, $p = 0.082$
	Study 4	47.3%	45.6%	$b = -0.018$, SE = 0.015, $p = 0.241$
NBA	Study 5	54.8%	50.4%	$b = -0.045$, SE = 0.018, $p = 0.014$
% of observations for which the second guess was in the right direction				
Preferences	Study 1	55.2%	49.6%	$b = -0.056$, SE = 0.018, $p = 0.002$
	Study 2	61.7%	53.7%	$b = -0.080$, SE = 0.016, $p < 0.001$
	Study 3	55.6%	53.1%	$b = -0.025$, SE = 0.010, $p = 0.015$
	Study 4	52.0%	48.6%	$b = -0.034$, SE = 0.016, $p = 0.031$
NBA	Study 5	64.1%	57.4%	$b = -0.067$, SE = 0.018, $p < 0.001$
% of observations for which the second guess was more extreme in the same direction than the first guess				
Preferences	Study 1	40.7%	48.6%	$b = 0.079$, SE = 0.021, $p < 0.001$
	Study 2	29.0%	40.4%	$b = 0.114$, SE = 0.019, $p < 0.001$
	Study 3	39.4%	45.9%	$b = 0.065$, SE = 0.012, $p < 0.001$
	Study 4	45.2%	49.9%	$b = 0.047$, SE = 0.020, $p = 0.017$
NBA	Study 5	29.9%	44.2%	$b = 0.143$, SE = 0.021, $p < 0.001$
Average absolute difference between first guess and second guess				
Preferences	Study 1	11.33	8.34	$b = -2.996$, SE = 0.589, $p < 0.001$
	Study 2	14.20	10.35	$b = -3.845$, SE = 0.538, $p < 0.001$
	Study 3	10.43	9.08	$b = -1.370$, SE = 0.308, $p < 0.001$
	Study 4	10.99	8.86	$b = -2.129$, SE = 0.646, $p = 0.001$
NBA	Study 5	7.47	6.74	$b = -0.731$, SE = 0.500, $p = 0.144$
Average within-participant correlation between first guess and second guess				
Preferences	Study 1	0.73	0.84	$b = 0.107$, SE = 0.029, $p < 0.001$
	Study 2	0.55	0.72	$b = 0.171$, SE = 0.033, $p < 0.001$
	Study 3	0.73	0.81	$b = 0.073$, SE = 0.017, $p < 0.001$
	Study 4	0.71	0.81	$b = 0.100$, SE = 0.033, $p = 0.003$
NBA	Study 5	0.47	0.60	$b = 0.120$, SE = 0.041, $p = 0.004$

Notes. The results for Studies 1, 2, 4, and 5 come from regressing the dependent measure on the Too High/Too Low condition, and the results for Study 3 come from regressing the dependent measures on (1) the Too High/Too Low condition, (2) the Second Guess Bonus condition, and (3) their interaction. The regression analyses included fixed effects for item (preference/behavior or game) and clustered standard errors by participant. Positive (negative) coefficients indicate that the Too High/Too Low question increased (decreased) the respective dependent measure. SE, standard error.

significant in all five studies. Thus, explicitly deciding in which direction to adjust their second guesses made people more likely to adjust in the wrong direction.

The fact that these studies differed in their participant populations (i.e., MTurk and the laboratory), incentive schemes,[12] and prediction tasks (i.e., forecasting others' preferences/behaviors and forecasting basketball game outcomes) attests to the generalizability of these effects. Thus, taken together, these studies provide strong evidence that explicitly deciding in which direction to adjust one's second guess makes people less likely to adjust in the right direction and renders the inner crowd less wise.

To understand *why* explicitly deciding in which direction to adjust one's first guess would decrease the probability of adjusting one's first guess in the right direction, we examined whether there were condition differences in the tendency to make a second guess that was more extreme than the first guess. In the first few studies that we conducted (Studies 1, 2, and 5), we noticed that most participants in the Estimate Only condition tended to retreat back toward the scale midpoint when making their second guesses. This could be a byproduct of a resampling process. For example, if you run the simulations of the resampling process that we report in the e-companion and then compute the percentage of the time that participants who use that process would make a second guess that was more extreme than their first guess, you find that they do so only 30%, 37%, and 43% of the time, respectively. That is, users of a resampling process will naturally adjust their first guesses in a less extreme

direction most of the time.[13] Users of a choice process, in contrast, may choose to adjust their first guess in a random direction if they don't know in which direction to adjust. If they do this, then their second guess would be more extreme relative to their first guess about half of the time. Or they may be more likely than not to make even more extreme second guesses if the quantity under consideration seems extreme to them (Simmons and Nelson 2006, 2020). If they do either of these things, then we would expect them to be more likely to adjust in an extreme direction than those who are using a resampling process.

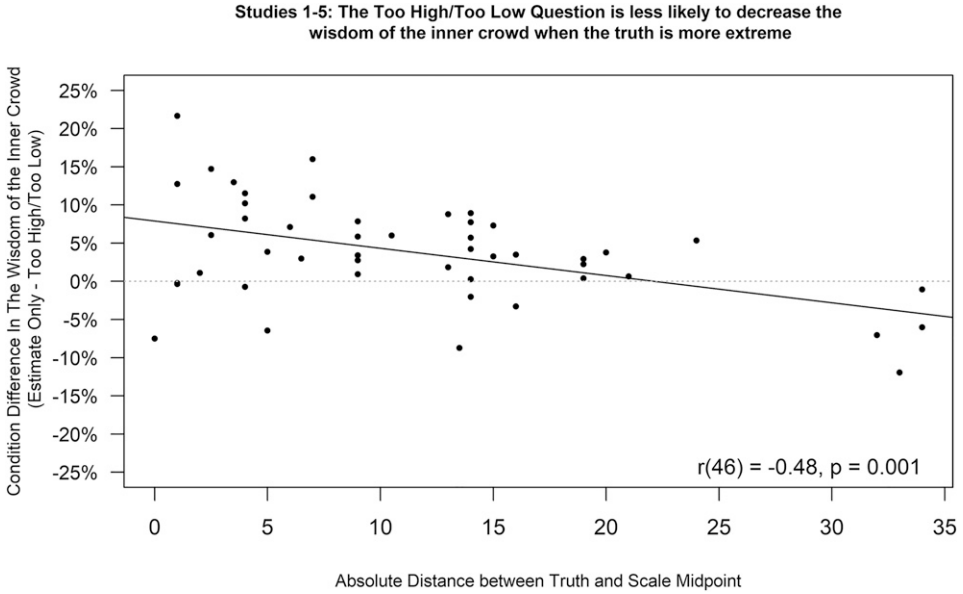
Thus, if the Too High/Too Low condition induced participants to engage in a choice process, then those who were asked to answer the Too High/Too Low question should have been more likely to make more extreme second guesses compared with those in the Estimate Only condition. As shown in Table 2, this is what we find. In Studies 1–5, participants in the Too High/Too Low condition made more extreme second guesses 49%, 40%, 46%, 50%, and 44% of the time, whereas those in the Estimate Only condition made more extreme second guesses only 41%, 29%, 39%, 45%, and 30% of the time, respectively. Indeed, in all five studies, we see a highly significant condition difference on this measure, one that is much larger than the effects of our manipulation on the likelihood of adjusting in the right direction and on the likelihood of the average guess outperforming the first guess. This suggested to us that answering the Too High/Too Low question may have reduced the wisdom of the inner crowd precisely because it made participants more likely to make second guesses that are more extreme in the same direction than their first guesses.[14] Indeed, in Studies 1–5, participants' first guesses were already too extreme relative to the scale midpoint 74%–78% of the time,[15] and therefore a process that induced even more extreme second guesses would often lead to second guesses that were in the wrong direction, rendering the inner crowd less wise.

Differentiating Between Two Alternative Explanations. Although the results of Studies 1–5 led us to believe that explicitly considering the direction of adjustment made the inner crowd less wise because it induced a choice process that altered the direction of adjustment, another explanation seemed viable. As shown in Table 2, asking participants to answer the Too High/Too Low question not only decreased the wisdom of the inner crowd, but it also (1) decreased the distance between participants' first and second guesses and (2) increased the within-subject correlation between participants' first and second guesses. Thus, one might wonder whether the Too High/Too Low question reduces the wisdom of the inner crowd by causing second guesses to be more extreme or by causing second guesses to be closer to first guesses. We can distinguish between these two explanations by focusing on cases in which people's first guesses tend to be too moderate rather than too extreme. If the Too High/Too Low question reduces the wisdom of the inner crowd by increasing the extremity of participants' second guesses, then it should not decrease the wisdom of the inner crowd when it is wise to adjust in a more extreme direction (i.e., when participants' first guesses are too moderate). If the Too High/Too Low question reduces the wisdom of the inner crowd simply by decreasing the distance between first and second guesses, then it should persist in doing so regardless of whether participants' first guesses are too moderate or too extreme.

In Studies 1–5, most of the correct answers were close to the midpoint of the scale. In these instances, we would expect most of people's estimates to be too extreme rather than too moderate, and, as reported previously, we found that participants' first guesses were too extreme the vast majority of the time. However, what if the true answers are extreme? We know that people tend to underestimate extremely large quantities and that they tend to overestimate extremely small quantities (Moore and Small 2007). As a result, people's first guesses of extreme quantities are likely to be too moderate rather than too extreme. Thus, if the direct effect of our Too High/Too Low manipulation is to increase the tendency to make more extreme second guesses, we should find that this manipulation *increases* the wisdom of the inner crowd when the true answers are extreme and people's first guesses are thus too moderate.

Although *most* of the questions in Studies 1–5 had moderate answers, not all of them did. To see whether the Too High/Too Low question was more likely to decrease the wisdom of the inner crowd for questions with moderate than extreme true answers, we investigated the across-item relationship between (1) how many units the true answer was away from the scale midpoint and (2) the difference in the percentage of the time the average guess was better than the first guess between the Estimate Only condition and the Too High/Too Low condition. As shown in Figure 3, there was a fairly strong relationship between these variables, such that the Too High/Too Low question was more likely to reduce the wisdom of the inner crowd for questions that had moderate rather than extreme true answers: $r(48) = -0.48$, $p = 0.001$.[16] Indeed, for the four questions with fairly extreme true answers, the Too High/Too Low question directionally *increased* the wisdom of the inner crowd. Although this is supportive of the notion that the Too High/Too Low question works to decrease the wisdom of the inner crowd by increasing the extremity of participants' second guesses, this analysis was

Figure 3. Across-Item Relationship Between the Extremity of an Item's Correct Answer (i.e., the Absolute Distance Between the Truth and the Scale Midpoint) and the Condition Difference in the Wisdom of the Inner Crowd (i.e., in the Percentage of Observations for Which the Average Guess Outperformed the First Guess)



Notes. This analysis was exploratory. R code to reproduce this figure: https://researchbox.org/45.

exploratory, and we did not design Studies 1–5 to look for it. That was what Study 6 was for.

Study 6: Moderate vs. Extreme True Answers

In Study 6, we tested how inducing a choice process affects the wisdom of the inner crowd for questions that have extreme rather than moderate true answers. As in Studies 1–4, we again asked participants to predict the percentage of survey responders who would report having a certain preference or engaging in a particular behavior. However, in this study, half of the prediction questions had moderate true answers and half had very extreme true answers.

Method

Participants. We conducted Study 6 on MTurk. Participants received $1 for participation, and they could also earn up to an additional $1 for accurate prediction performance. We decided in advance to collect data from 1,300 participants, and, after our preregistered exclusions, we wound up with a final sample of 1,280 participants (58% female, average age = 34 years). Our final sample included all participants who made a first and second guess for at least one of the prediction questions, and, as preregistered, excluded those whose IP addresses appeared twice in the same data set (10 participants) and whose IP addresses matched IP addresses in data sets from the previous studies we conducted on this topic (24 participants).

Procedure and Dependent Measures. The procedure of Study 6 was identical to that of Study 1 except for two changes. First, we used a modified set of prediction questions that allowed us to manipulate the extremity of the true answers within subjects (Table 3). We included six questions with moderate true answers (i.e., between 30% and 70%) and six questions with extreme true answers (i.e., below 20% or above 80%). We selected the questions with moderate true answers from Study 1, and we selected the new questions with extreme true answers based on a pretest. Second, as in Study 2, we incentivized participants by providing a bonus of $1 if the better of participants' guesses were on average less than 10 percentage points away from the true answers. We preregistered to analyze the same five measures that we analyzed in Studies 1–5 (Table 2).

Results and Discussion

Analysis Plan. In the course of conducting our analyses, we realized that the survey contained an error for the vanilla versus chocolate ice cream item. When asked for their first guess, participants were asked to estimate the percentage of people who preferred vanilla to chocolate ice cream, but when asked for their second guess, participants were asked to estimate the percentage of people who preferred chocolate to vanilla ice cream (and were erroneously told that their first estimate had been about the percentage of people who preferred chocolate to vanilla ice cream).

Table 3. Prediction Questions and True Answers in Study 6

Extremity of true answer	Percentage of participants who	Study 6
Moderate	… prefer having a cat to having a dog	36%
	… own an iPad	42%
	… drink at least 10 cups of coffee in a week	42%
	… prefer vanilla ice cream to chocolate ice cream	49%
	… have posted a video on YouTube	54%
	… prefer Spring to Summer	68%
Extreme	… have tried bungee jumping	5%
	… have swam with sharks	6%
	… have traveled outside the United States within the last month	9%
	… prefer having a cat to having a chinchilla	81%
	… prefer chocolate ice cream to cheese ice cream	94%
	… have a TV	95%

Because of this, we deleted this item from the analysis. Including it does not affect our conclusions.

After removal of this item, each participant who fully completed the study contributed 11 observations to the data set, one for each of the preferences/behaviors they estimated. We dropped 37 observations from the analyses because the second guess was missing. We regressed the dependent measures on the (1) Too High/Too Low condition (−0.5 = Estimate Only; +0.5 = Too High/Too Low), (2) the Extreme True Answer condition (−0.5 = Moderate; 0.5 = Extreme), and (3) their interaction. We present the results from OLS regressions; logistic regressions yielded nearly identical results. All our regressions clustered standard errors by participant.[17]

Main Analysis. Table 4 presents the descriptive statistics and the results of the significance tests. As predicted, the Too High/Too Low question had different effects on the wisdom of the inner crowd depending on whether the prediction questions had moderate or extreme true answers. For questions with moderate true answers, the Too High/Too Low question decreased the likelihood that the average guess would outperform the first guess ($p = 0.021$) and also decreased the likelihood that participants' second guesses would be in the right direction relative to their first guesses ($p = 0.002$). This is consistent with what we found in Studies 1–5. However, for questions with extreme true answers, we observed the opposite: the Too High/Too Low Question *increased* the likelihood that the average guess would outperform the first guess ($p < 0.001$) and also marginally increased the likelihood that participants' second guesses would be in the right direction relative to their first guesses ($p = 0.084$). Thus, it seems that answering the Too High/Too Low question decreases the wisdom of the inner crowd only for questions with moderate true answers and not for questions with extreme true answers.

The effects of the Too High/Too Low question on the extremity of participants' second guesses were interesting. As in Studies 1–5, we found that, overall, participants were much more likely to give a more extreme second guess in the Too High/Too Low condition than in the Estimate Only condition ($p < 0.001$). However, we also found that this effect was much stronger for questions with extreme true answers than for questions with moderate true answers ($p < 0.001$). This result suggests that participants who are using the choice process (i.e., those in the Too High/Too Low condition) do not always respond randomly when considering the direction of adjustment. Instead, they are more likely to give more extreme second guesses when the quantity they are considering is extreme. This result is consistent with previous work showing that people are more likely to predict that their prior estimates of extreme quantities were too moderate (Simmons and Nelson 2006, 2020).

Finally, analyses of the average absolute deviations between guesses and the average within-subject correlation again showed that participants in the Too High/Too Low condition tended to provide guesses that were closer together and that were more highly correlated with each other (see Table 4).

In sum, the results of Study 6 strongly suggest that the direct effect of the Too High/Too Low question is to induce participants to engage in a choice process that results in them giving more extreme second guesses than they otherwise would have. When the true answers are moderate, this harms the wisdom of the inner crowd. However, when the true answers are extreme, this improves the wisdom of the inner crowd.[18]

Study 7: What About A Different Too High/Too Low Question?

In Study 7, we investigated the specificity and robustness of our findings. We again asked participants to predict the percentage of survey responders who

Table 4. Results for Study 6

	Estimate Only	Too High/ Too Low	Effect of Too High/Too Low condition (vs. Estimate Only condition)	Effect of Extreme True Answer condition (vs. Moderate True Answer condition)	Interaction
% of observations for which the average guess outperformed the first guess					
All data	52.1%	53.4%	$b = 0.009$, SE = 0.009, $p = 0.304$	$b = 0.061$, SE = 0.009, $p < 0.001$	$b = 0.078$, SE = 0.019, $p < 0.001$
Moderate true answer	50.9%	47.9%	$b = -0.030$, SE = 0.013, $p = 0.021$		
Extreme true answer	53.1%	57.9%	$b = 0.048$, SE = 0.013, $p < 0.001$		
% of observations for which the second guess was in the right direction					
All data	58.3%	57.7%	$b = -0.009$, SE = 0.009, $p = 0.275$	$b = 0.082$, SE = 0.009, $p < 0.001$	$b = 0.062$, SE = 0.019, $p = 0.001$
Moderate true answer	55.6%	51.5%	$b = -0.041$, SE = 0.013, $p = 0.002$		
Extreme true answer	60.6%	62.8%	$b = 0.022$, SE = 0.013, $p = 0.084$		
% of observations for which the second guess was more extreme in the same direction than the first guess					
All data	43.2%	53.2%	$b = 0.096$, SE = 0.012, $p < 0.001$	$b = 0.079$, SE = 0.009, $p < 0.001$	$b = 0.086$, SE = 0.017, $p < 0.001$
Moderate true answer	41.3%	46.6%	$b = 0.053$, SE = 0.014, $p < 0.001$		
Extreme true answer	44.8%	58.7%	$b = 0.139$, SE = 0.015, $p < 0.001$		
Average absolute difference between first guess and second guess					
All data	10.14	8.29	$b = -1.854.$, SE = 0.322, $p < 0.001$	$b = -2.134$, SE = 0.147, $p < 0.001$	$b = 0.008$, SE = 0.294, $p = 0.978$
Moderate true answer	11.31	9.45	$b = -1.858$, SE = 0.329, $p < 0.001$		
Extreme true answer	9.17	7.32	$b = -1.850$, SE = 0.378, $p < 0.001$		
Average within-participant correlation between first guess and second guess					
All data	0.82	0.88	$b = 0.058$, SE = 0.013, $p < 0.001$	$b = 0.189$, SE = 0.009, $p < 0.001$	$b = -0.088$, SE = 0.018, $p < 0.001$
Moderate true answer	0.70	0.80	$b = 0.101$, SE = 0.019, $p < 0.001$		
Extreme true answer	0.94	0.95	$b = 0.014$, SE = 0.011, $p = 0.200$		

Notes. The results for Study 6 come from regressing the dependent measure on (1) the Too High/Too Low condition, (2) the Extreme True Answer condition, and (3) their interaction in analyses that clustered standard errors by participant. Positive (negative) coefficients indicate that the Too High/Too Low question increased (decreased) the respective dependent measure.

would report having a certain preference or engaging in a particular behavior. However, in this study, we added an additional condition in which we asked participants to answer a new version of the Too High/Too Low question before they made their second guess. Specifically, we asked participants to decide whether 50% is too high or too low. Adding this new condition allowed us to test whether our results were specific to people explicitly considering in which direction their first guess was wrong, or whether they emerge when people are asked a different type of Too High/Too Low question. In addition, in our previous studies with preference/behavior questions, participants in all conditions were always asked to indicate whether they thought the majority of survey responders had the preference/behavior before making a guess. It is possible that the inclusion of this question influenced our results, and so in Study 7 we omitted it.

Method

Participants. We conducted Study 7 on MTurk. Participants received $1 for participation. They could also earn up to an additional $1 for accurate prediction performance. We decided in advance to collect data from 1,000 participants, and, after our preregistered exclusions, we wound up with a final sample of 972 participants (average age = 36 years; 59% female). Our final sample included all participants who made a first and second guess for at least one of the prediction questions, and, as preregistered, excluded those with IP addresses that appeared twice in the same data set (8 participants), and whose IP addresses matched IP addresses in data sets from the previous studies we conducted on this topic (22 participants).

Procedure and Dependent Measures. Study 7 followed the same procedure as Study 1 except for three changes. First, in contrast to our previous studies investigating preference/behavior questions, participants were not asked to indicate whether they thought the majority of survey responders exhibited the preference/behavior prior to estimating the percentage. Second, participants were randomly assigned to one of three conditions. The first two conditions were those that we had previously tested in Studies 1–6: For each prediction question, before making their second guess, we either asked participants to answer the Too High/Too Low question or not. In our new third condition, we asked participants to answer a 50% Too High/Too Low question. This question asked participants to indicate whether they thought that more or fewer than 50% of survey responders would report having the preference or engaging in the behavior (e.g., "Do you think that more or fewer than 50 percent of survey responders are more interested in politics than sports?"). Thus, this condition was similar to the original Too High/Too Low condition, but we asked whether the true answer was above or below 50% rather than above or below the participant's first guess. Finally, in this study, we used the same bonus payment rule as in Studies 2 and 6. That is, we incentivized participants by providing a bonus of $1 if the better of participants' two guesses was on average less than 10 percentage points away from the true answers. We preregistered to analyze the same five measures as in the previous study.

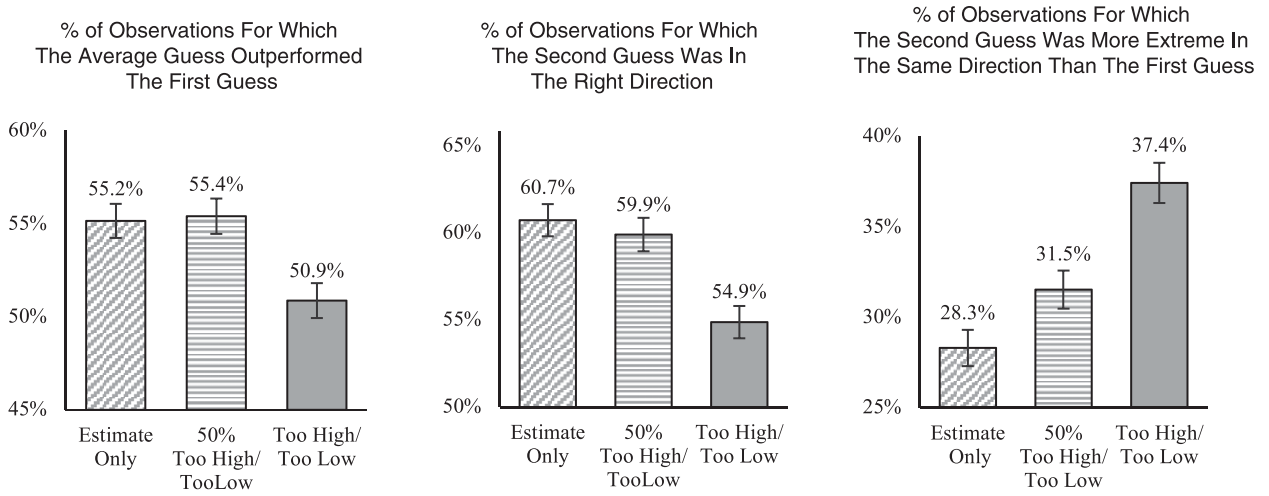
Results and Discussion

Analysis Plan. Each participant who fully completed the study contributed 10 observations to the data set, one for each of the preferences/behaviors they estimated. We conducted two regression analyses for each of our dependent measures. First, we regressed the dependent measures on (1) the Too High/Too Low condition and (2) the 50% Too High/Too Low condition. This allowed us to compare these two conditions to the Estimate Only condition. Second, we regressed the dependent measures on (1) the Estimate Only condition and (2) the 50% Too High/Too Low condition. This allowed us to compare these two conditions to the Too High/Too Low condition. All our regressions included fixed effects for prediction question and clustered standard errors by participant.

Main Analysis. We present the results from Study 7 in Figure 4. First, comparing the Too High/Too Low condition to the Estimate Only condition, we replicated our previous findings using questions with moderate true answers: Explicitly answering the Too High/Too Low question decreased the likelihood that the average guess would outperform the first guess, $b = -0.057$, SE = 0.017, $t(971) = -3.27$, $p = 0.001$, reduced the likelihood that participants' second guesses would be in the right direction relative to their first guesses, $b = -0.078$, SE = 0.017, $t(971) = -4.47$, $p < 0.001$, and increased the likelihood that participants' second guesses would be more extreme relative to their first guesses, $b = 0.122$, SE = 0.020, $t(971) = 6.10$, $p < 0.001$.

Importantly, however, for all our three primary dependent measures, the effects of the Too High/Too Low condition also significantly differed from the effects of the new 50% Too High/Too Low condition. Compared with explicitly answering the 50% Too High/Too Low question, explicitly answering the Too High/Too Low question decreased the likelihood that the average guess would outperform the first guess, $b = -0.060$, SE = 0.018, $t(971) = -3.40$, $p = 0.001$, decreased the likelihood that participants' second guesses would be in the right direction relative to their first guesses, $b = -0.067$, SE = 0.018, $t(971) = -3.77$, $p < 0.001$, and also increased the likelihood that participants' second guesses would be more extreme

Figure 4. Results of Study 7



relative to their first guesses, $b = 0.079$, SE = 0.020, $t(971) = 3.86$, $p < 0.001$. This suggests that the effects we observed in Studies 1–6 are not produced by *any* Too High/Too Low question; rather they seem specific to considering whether one's first guess was too high or too low.

The 50% Too High/Too Low condition and the Estimate Only condition did not differ with respect to their effects on the percentage of observations for which the average guess outperformed the first guess ($p = 0.849$) or for which the second guess was in the right direction relative to the first guess ($p = 0.536$). However, compared with the Estimate Only condition, explicitly answering the 50% Too High/Too Low question increased the likelihood that participants' second guesses would be more extreme relative to their first guesses, $b = 0.043$, SE = 0.019, $t(971) = 2.22$, $p = 0.027$.

Finally, the results of the absolute deviation and correlation measures provide further evidence that our effects are not driven by the difference in the independence of guesses between the Too High/Too Low and Estimate Only conditions. Although we again found that compared with participants in the Estimate Only condition, participants in the Too High/Too Low condition were more likely to provide guesses that were closer together, $b = -3.610$, SE = 0.611, $t(971) = -5.91$, $p < 0.001$, and more highly correlated with each other, $b = 0.159$, SE = 0.031, $t(968) = 5.11$, $p < 0.001$, we also found this to be true for participants in the 50% Too High/Too Low condition. That is, compared with participants in the Estimate Only condition, participants in the 50% Too High/Too Low condition were also more likely to provide guesses that were closer together, $b = -2.995$, SE = 0.614, $t(971) = -4.88$, $p < 0.001$, and more highly correlated with each other, $b = 0.144$, SE = 0.032, $t(968) = 4.49$, $p < 0.001$. The Too High/Too Low condition and the 50% Too High/Too Low condition did not differ from each other with respect to these two measures ($p = 0.277$ and $p = 0.591$, respectively). If increasing the correlation between first and second guesses necessarily decreases the wisdom of the inner crowd, then the 50% Too High/Too Low condition should have exhibited a smaller wisdom-of-the-inner-crowd effect than the Estimate Only condition. However, it did not.

In sum, Study 7 suggests that our results are specific to asking participants to decide whether their first guess is too high or too low and do not occur when participants are asked to decide whether 50% is too high or too low. An additional study, Study S1 in the e-companion, provides further evidence for this point. In this study, we tested a new Too High/Too Low question: Before participants made their second guess, we asked them to indicate whether they thought that a given answer to a different estimation question (i.e., an estimation question other than the question for which they made their first guess) was greater or less than the true percentage. This did not influence the wisdom of the inner crowd, again suggesting that the effects documented herein are not caused merely by the introduction of any sort of question.

Study 8: Unbounded Scales

In Studies 1–7, we found that asking people to decide whether their first guess was too high or too low leads them to make more extreme second guesses. When the correct answers are not extreme, this intervention reduces people's tendency to provide a second guess that is in the right direction and harms the benefits of averaging, thus rendering the inner crowd less wise.

In most of the studies we have conducted thus far, the questions that we posed to participants were on a bounded scale, with endpoints at 0%–100%. Although many important predictions are made on these kinds of bounded scales (e.g., probability forecasts), it is important to know whether our effects generalize to questions that are answered on unbounded scales.

Although the results of Study 5—in which we asked people to predict the outcomes of NBA basketball games on an unbounded scale (points)—suggest that it does, we wanted to further establish this by investigating predictions in another domain: stocks.

In Study 8, we asked participants to predict the stock prices of 10 companies two weeks in advance. As in the previous studies, they made both first and second guesses, and we manipulated whether they were asked whether their first guess was too high or too low before making their second guess. Unlike in previous studies, the domain of stock predictions does not make it easy to assess whether any given prediction is wise, as short-term stock market changes are largely random. Because of this, in Study 8 we focused on establishing that our manipulation changes participants' second guesses in the same manner as in our previous studies rather than on investigating condition differences in the wisdom of the inner crowd. Specifically, we focused on examining whether the Too High/Too Low question would increase participants' tendency to give a second guess that was in a more extreme direction relative to their first guess. We preregistered this to be our primary dependent measure in this study.

Of course, we could try to examine differences in the wisdom of the inner crowd in this study, by comparing participants' predictions of stock price changes to how those prices actually changed. Although we caution against putting too much faith in this noisy measure (or against relying on it in replications), we preregistered it as a secondary analysis and we do report it below. The results are consistent with those reported in our previous studies.

Method

Participants. We conducted Study 8 on MTurk. Participants received $1.20 for participation. They could also earn an additional $1 for accurate prediction performance. We decided in advance to collect data from 1,200 participants, and we wound up with a final sample of 1,191 participants (average age = 39.8 years; 48% female). Our final sample included all participants who made a first and second guess for at least one of the prediction questions, and, as preregistered, excluded those whose IP addresses appeared twice in the same data set (19 participants).

Procedure and Dependent Measures. In this study, participants predicted the stock prices of 10 well-known companies (Apple, Comcast, Microsoft, Netflix, PayPal, Pepsico, Tesla, T-Mobile, United, and Walgreens). For each stock, participants saw the name of the company, the stock symbol, and the current price of the stock. We posted the study on Saturday, March 16, 2019, and so the current price provided to participants was that as of market close on Friday, March 15, 2019 (see Table S8 in the e-companion).

For each stock, participants were asked to predict what the stock price would be at the end of the day on Friday, March 29, 2019 (two weeks later). Before participants made their prediction, we first asked them to indicate whether they thought that the stock price would be (a) higher than the current stock price or (b) lower than or equal to the current stock price. For example, the stock price of Microsoft at the end of the day on Friday, March 15, 2019 was $115.91, and so we asked, "Do you think the stock price of Microsoft will be higher or lower than $115.91 when the stock market closes on March 29, 2019?" We included this question to ensure that participants would pay attention to the current stock price. After answering this question, participants made their own prediction. The survey software required participants to enter a price with two decimals when making their prediction, and we presented each stock on a separate page in a random order. Participants were incentivized to make accurate predictions. They received $1 if, on average across all stocks, their predictions were within 1% of the true stock prices.

After making their predictions for all 10 stocks, participants were (unexpectedly) asked to predict the stock prices again. They were asked "to make a second prediction that is not exactly the same as your first prediction," and they were told that making a second guess would give them "a second chance to make accurate predictions and thus a second chance to earn a bonus." Participants were presented with each of the stocks again (in random order). For each stock, participants were shown their first guess along with their prediction of whether their first guess would be higher or lower than the current stock price (e.g., "You predicted that the price of this stock will be [higher than/lower than or equal to] $115.91 when the stock market closes on March 29, 2019. You predicted that it would be [first guess].") Participants were then asked to make a second different guess.

Similar to Studies 1–6, we manipulated (between subjects) whether we asked participants to decide whether their first guess was too high or too low right before they made their second guess. Participants in the Too High/Too Low condition were asked to decide whether their first guess was too high or too low (e.g., "Do you think the stock price of Microsoft will be higher or lower than [first guess] when the stock market closes on March 29, 2019?"). Participants in the Estimate Only condition were not asked the Too High/Too Low question but were only asked to make a second guess.[19]

Results and Discussion

Analysis Plan. We focus our analysis in this study on the extremity measure: the binary variable indicating,

for each observation, whether the participant's second guess was more extreme in the same direction than her first guess. We define extremity relative to the current price of the stock. That is, a participant's second guess was considered to be more extreme in the same direction than her first guess if it was further away from the current price of the stock in the same direction than her first guess was. For example, if the current price of the stock was \$115.91 and a participant's first guess was \$116.50, then any second guess above \$116.50 would be considered a more extreme second guess, and any second guess less than \$116.50 would be considered a more moderate second guess.

Each participant who fully completed the study contributed 10 observations to the data set, one for each of the stocks they predicted. We dropped 25 observations because the first and second were identical and 56 observations because the participant's second guess was missing, preventing us from being able to calculate our dependent measures.[20] We regressed whether the participant's second guess was more extreme than her first guess on the Too High/Too Low condition, including fixed effects for stocks and clustering standard errors by participant. For completeness, we also report the results from the right-direction and accuracy measures below using the same analytic strategy.

Main Analysis. We present the results of Study 8 in Table 5. Replicating the results from our previous studies, we found that answering the Too High/Too Low question increased the likelihood that participants' second guesses would be more extreme relative to their first guesses, $b = 0.038$, SE = 0.017, $t(1190) = 2.22$, $p = 0.027$. Thus, we find that the effects of the Too High/Too Low question hold in another prediction domain that has an unbounded response scale.

We did not predict strong effects on our measures of accuracy given how noisy stock prices are. Nevertheless, we saw that answering the Two High/Too Low question decreased the likelihood that participants' second guesses would be in the right direction relative to their first guesses, $b = -0.023$, SE = 0.011, $t(1190) = -2.15$, $p = 0.032$, and it directionally, but not significantly, decreased the percentage of observations for which the average guess outperformed the first guess, $b = -0.017$, SE = 0.011, $t(1190) = -1.61$, $p = 0.108$.

Although we did not preregister any additional measures, we can also see whether participants' guesses were closer together and more highly correlated when they were asked to answer the Too High/Too Low question, as was the case in the previous studies. Given that we used a range of different stock prices for this study—ranging from as low as \$40.47 (Comcast) to as high as \$361.46 (Netflix)—we converted each participants' predictions into their predicted percentage change with respect to the current stock price. We also removed extreme observations for which participants predicted stock price changes of more than 50%, resulting in a total of 130 exclusions (1.1% of all observations).[21] After these exclusions, we did not observe any significant differences between the conditions with respect to the absolute difference between guesses ($p = 0.888$), but we again observed

Table 5. Results for Study 8

Estimate Only	Too High/Too Low	Effect of the Too High/Too Low condition (vs. Estimate Only condition)
% of observations for which the average guess outperformed the first guess		
49.1%	47.3%	$b = -0.017$, SE = 0.011, $p = 0.108$
% of observations for which the second guess was in the right direction		
54.2%	51.9%	$b = -0.023$, SE = 0.011, $p = 0.032$
% of observations for which the second guess was more extreme in the same direction than the first guess		
44.0%	47.8%	$b = 0.038$, SE = 0.017, $p = 0.027$
Average absolute difference between first guess and second guess[a]		
2.29	2.27	$b = -0.022$, SE = 0.157, $p = 0.888$
Average within-participant correlation between first guess and second guess		
0.50	0.60	$b = 0.098$, SE = 0.031, $p = 0.001$

Notes. The results come from regressing the dependent measure on the Too High/Too Low condition, including fixed effects for stock and clustered standard errors by participant. Positive (negative) coefficients indicate that the Too High/Too Low question increased (decreased) the respective dependent measure.

[a]To compute these means, we first converted all guesses to percent changes from the current stock price.

that participants' first and second guesses were more highly correlated when they first answered the Too High/Too Low question ($p = 0.001$).

Taken together, the results from Study 8 demonstrate that answering the Too High/Too Low question alters participants' second guesses in the domain of stock prices, just as it does in the domains we investigated in Studies 1–7.

General Discussion

Past research suggests that people can improve their quantitative judgments by making two judgments and then taking the average, an effect called *the wisdom of the inner crowd* (Ariely et al. 2000; Vul and Pashler 2008; Herzog and Hertwig 2009, 2014a b). We discovered that this effect hinges on whether, while making their second guesses, people consciously consider the direction in which their first guess had erred.

We conducted nine experiments in which people provided two guesses of some quantity. Before providing their second guesses, all participants were reminded of their first guess, and we manipulated whether participants were explicitly asked to indicate whether that guess was too high or too low before they made their second guess. We found that asking people to explicitly answer this "Too High/Too Low question" before making their second guesses decreased people's tendency to provide second guesses that were in the right direction and reduced (and sometimes eliminated or reversed) the wisdom-of-the-inner-crowd effect. Thus, people are usually more likely to give second guesses that are in the right direction when they do not explicitly decide in which direction their first guess had erred.

We also found that these effects are explained by the fact that the introduction of the Too High/Too Low question alters the direction of people's second guesses. Specifically, people who are asked the Too High/Too Low question are more likely to give second guesses that are more extreme (in the same direction) than those who are not asked this question. As demonstrated in Study 6, this means that the Too High/Too Low question will reduce the wisdom of the inner crowd when people's first guesses are too extreme (as they usually are when true answers are moderate) but will enhance the wisdom of the inner crowd when people's first guesses are too moderate (as they usually are when true answers are extreme). In Study 6, we also found that asking the Too High/Too Low question was even more likely to induce extreme second guesses when the true answers were extreme than when they were moderate.

Based on these results, we believe that explicitly considering the direction in which one's first guess had erred increases people's tendency to use that first guess as a starting point from which their second guesses are generated. For quantities that are not extreme, this may increase people's tendency to adjust in a random direction, particularly if they realize that they do not know in which direction to adjust. For quantities that are extreme, we expect that people answer the question of whether their first guess is too high or too low by relying on intuitions about the quantity under consideration. Research by Simmons and Nelson (2006, 2020) shows that people often answer questions of the form, "Is X greater or less than Y?" by relying on an intuition that is based on the simpler evaluation, "Is X large or small?" When X is a moderately sized quantity, this process may lead participants to respond randomly; but when X is extreme, it leads to predictable choices. For example, when X is extremely large, people tend to judge it to be larger than Y, so long as Y itself is not obviously too large or too small. Applied to the current research, this means that people who are explicitly asked whether the percentage of responders who own a television is higher or lower than their first guess of 91% would have an (in this case correct) intuition that it is higher, and their second guess would be more extreme.

It is important to emphasize one limitation of our investigation. Here we have shown that explicitly deciding in which direction one's first guess was wrong increases people's tendency to produce second guesses that are in a more extreme direction relative to their first guesses. That is, they are less likely to revert back toward the midpoint of the scale when providing their second guesses. For the many important questions that have explicit midpoints (like 50%) or natural midpoints (like the tie game midpoint in Study 5 or the current stock price midpoint in Study 8 or any forecast made on a percentage scale), our investigation provides generalizable and replicable insights. Nevertheless, some questions that people are asked to forecast do not offer easy-to-identify scale midpoints (e.g., how many total points will be scored in this game?), and in such cases it will be hard to predict what effect our Too High/Too Low manipulation will have on the wisdom of the inner crowd. To do so, researchers will have to find a way to reliably assess participants' psychological midpoints.

Our results contribute to a growing literature on the wisdom-of-the-inner-crowd effect (Ariely et al. 2000; Vul and Pashler 2008; Herzog and Hertwig 2009, 2014a, b; van Dolder and van den Assem 2018). We make at least two contributions.

First, because previous researchers have found evidence for the wisdom of the inner crowd, they have concluded that "responses made by a subject are sampled from an internal probability distribution" (Vul and Pashler, 2008, p. 647), and that "subjects did not produce a second guess by simply perturbing the first" (p. 646). Our investigation suggests that the validity of

Table 6. Relative Accuracy of First and Second Guesses

Study	Items	% of time first guess was better than second guess	Test of difference from 50%
		Estimate Only condition	
Study 1		53.6%	$b = 0.036$, SE = 0.012, $p = 0.004$
Study 2		51.2%	$b = 0.012$, SE = 0.011, $p = 0.276$
Study 3		52.7%	$b = 0.027$, SE = 0.007, $p < 0.001$
Study 4		57.0%	$b = 0.070$, SE = 0.010, $p < 0.001$
Study 5		37.6%	$b = -0.124$, SE = 0.013, $p < 0.001$
Study 6	All data	52.7%	$b = 0.027$, SE = 0.007, $p < 0.001$
	Moderate items	53.6%	$b = 0.036$, SE = 0.009, $p < 0.001$
	Extreme items	51.9%	$b = 0.019$, SE = 0.010, $p = 0.044$
Study 7		49.9%	$b = -0.001$, SE = 0.009, $p = 0.880$
Study 8		54.3%	$b = 0.043$, SE = 0.008, $p < 0.001$
Study S1		51.0%	$b = 0.010$, SE = 0.007, $p = 0.129$
		Too High/Too Low condition	
Study 1		56.3%	$b = 0.063$, SE = 0.012, $p < 0.001$
Study 2		55.3%	$b = 0.053$, SE = 0.010, $p < 0.001$
Study 3		53.7%	$b = 0.037$, SE = 0.007, $p < 0.001$
Study 4		57.8%	$b = 0.078$, SE = 0.011, $p < 0.001$
Study 5		42.5%	$b = -0.075$, SE = 0.012, $p < 0.001$
Study 6	All data	50.9%	$b = 0.009$, SE = 0.006, $p = 0.141$
	Moderate items	56.1%	$b = 0.061$, SE = 0.009, $p < 0.001$
	Extreme items	46.5%	$b = -0.035$, SE = 0.010, $p < 0.001$
Study 7		53.2%	$b = 0.032$, SE = 0.010, $p = 0.001$
Study 8		56.0%	$b = 0.060$, SE = 0.008, $p < 0.001$
Study S1		54.1%	$b = 0.041$, SE = 0.006, $p < 0.001$

Notes. Each of these percentages reflects the percentage of observations for which first guesses were better than second guesses, after removing observations for which first and second guesses were equally accurate. The significant tests reflect the results of OLS regressions in which we regressed the difference between the percentage and 50% on a constant while clustering standard errors by participant. The key test in this regression is the intercept, which is what is reported in the final column.

these conclusions depends on how second guesses are generated. If people always use a resampling process to generate their second guesses, then the wisdom of the inner crowd will be robust and reliable. However, if some people, either by instruction or by natural inclination, explicitly consider the direction in which their first had erred prior to giving their second guess, then the wisdom of the inner crowd will be less robust and reliable.

Indeed, our results suggest that, even within the Estimate Only control condition, a resampling process does not fully govern how people generate their second guesses. If people generate their second guesses using a resampling process, then two things should be true. First, we should always observe a wisdom-of-the-inner-crowd effect. Second, we should observe that first guesses and second guesses are equally accurate.[22] However, we did not observe these things. For example, in Studies 4 and 8, there was no wisdom-of-the-inner-crowd effect within the Estimate Only condition. As shown in Table 6, in six of our nine studies, second guesses and first guesses exhibited significant differences in accuracy even among those in the Estimate Only condition, with first guesses being significantly better in five of our nine studies and significantly worse in one study (Study 5). Although we suspect that resampling accounts for some fraction of how participants' second guesses are naturally generated, it is clear that it cannot account for all of it. It is possible that at least some people spontaneously generate their second guess by explicitly considering the direction in which their first guess had erred.

Second, our finding that the wisdom of the inner crowd usually suffers when people consider whether their first guess was too high or too low is surprising in light of the findings of Herzog and Hertwig (2009, 2014b) that asking participants this question (without requiring them to answer it) as part of their *dialectical bootstrapping* manipulation actually increased the wisdom of the inner crowd. Indeed, there were many ex ante reasons to expect the introduction of the Too High/Too Low question to enhance the wisdom-of-the-inner-crowd effect. For example, it could have prompted participants to consider all the ways their first guess was wrong, eliciting a more different second guess. Or, it could have prompted participants to reflect more carefully on their first guess and to realize the nature of their error. Nothing in the prior literature would have led readers to predict that answering the Too High/Too Low question would lead people to give second guesses that were more

extreme (in the same direction) than their first guesses. We certainly did not.

This finding is not just surprising, but it is important, both theoretically and practically. Theoretically, it suggests that second guesses are not always generated by a resampling process, that the wisdom of the inner crowd is not inevitable, and that we cannot predict whether the wisdom-of-the-inner-crowd effect will emerge without understanding the processes people use to generate second guesses. Practically, it suggests that the wisdom of the inner crowd will be most likely to emerge when people are discouraged (or at least not encouraged) from considering in which direction their first guess was wrong. In general, managers who attempt to elicit the wisdom of the inner crowd from their employees should carefully design the elicitation of second guesses to discourage them from considering how their first guess was wrong, at least in the majority of cases in which the correct answers are not likely to be extreme.

Acknowledgments

The authors thank Uri Simonsohn, Deborah Small, and Joshua Lewis for helpful feedback and Sara Sermarini, David Eskenazi, Beidi Hu, and Soaham Bharti for excellent research assistance. Studies were conducted at the University of Chicago and the University of Pennsylvania.

Appendix. Links to Pre-Registrations

Study 1: https://aspredicted.org/393hr.pdf
Study 2: https://aspredicted.org/xv5ru.pdf
Study 3: https://aspredicted.org/w5fi6.pdf
Study 4: https://aspredicted.org/gc58f.pdf
Study 5: https://aspredicted.org/ae5ub.pdf
Study 6: https://aspredicted.org/9x25q.pdf
Study 7: https://aspredicted.org/q6gf4.pdf
Study 8: https://aspredicted.org/kh2yg.pdf
Study S1: https://aspredicted.org/36ee8.pdf

Table A1. Content of the e-Companion

Section	Pages
Supplement 1: Results from Simulations	1–12
Supplement 2: Results Without Excluding Duplicate Laboratory IDs in Study 2	13
Supplement 3: Results for the Choice Only Condition in Studies 1 and 5	14
Supplement 4: Mediation Results for Studies 1–5	15–17
Supplement 5: Additional Results for Studies 1–8	18–19
Supplement 6: Stocks used in Study 8	20
Supplement 7: Study S1	21–24

Endnotes

[1] It may seem obvious that, under this process, a person's average guess will usually be better than her first guess if her guesses are drawn from a distribution that is centered on the truth. Probably less obvious is the fact that, under this process, a person's average guess will usually be better than her first guess even if her guesses are drawn from a distribution that is centered on a value that is pretty far from the correct answer (e.g., 2 standard deviations away from it). Indeed, the use of this kind of a resampling process will almost always produce a wisdom-of-the-inner-crowd effect, and it will never produce the opposite effect (see the simulations that we present in the e-companion).

[2] We obtained the dialectical bootstrapping instructions from the authors of Herzog and Hertwig (2014b). Those instructions were in German, and we translated them into English. Alternative wordings of these instructions as they are described in Herzog and Hertwig (2009, 2014b) read as follows: "First, assume that your first estimate is off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was the first estimate rather too high or too low? Fourth, based on this new perspective, make a second, alternative estimate" (Herzog and Hertwig 2014b, p. 221).

[3] As preregistered, in Studies 3 and 4, we excluded data from IP addresses that appeared twice in the same data set (21 participants in Study 3 and 5 participants in Study 4) or that matched IP addresses in data sets from previous studies (27 participants in Study 3 and 16 participants in Study 4).

[4] Because we thought that participants would be prevented from taking the study twice, we did not preregister to delete duplicate responders in this study. However, we think it makes sense to remove duplicate responses of participants who took our study more than once. Our results do not change if we do not exclude these participants. We present the results for the full data set in the e-companion.

[5] After participants made all their first and second guesses, we asked them two exploratory questions. First, we asked them which of their two sets of predictions they thought was more accurate: "If you had to bet, which set of your estimates do you think is more accurate on average, the first set or the second set?" Second, we asked them, "Did you try harder to be accurate when making your first set of estimates or when making your second set of estimates?" (1 = "The first set," 2 = "I was trying equally hard on the first set and the second set," 3 = "The second set"). We have not yet fully analyzed these exploratory measures, and we do not discuss them further.

[6] We chose not to analyze this condition in the main text or to include it in subsequent studies (Studies 1 and 5 were run before the others), because this condition differed in too many ways from the others. Most notably, people's second guesses were made many minutes after their answers to the (possibly forgotten) Too High/Too Low question and were made after answering the Too High/Too Low questions of all of the other items. Thus, it is not clear how to interpret any differences involving this condition. As shown in the e-companion, the results from this Choice Only condition tended to fall in between those of the Estimate Only and Too High/Too Low conditions. Our decision to exclude this condition from the main text means that the analyses reported for Studies 1 and 5 are slightly different from those that we preregistered, in the sense that we preregistered to compare three conditions rather than two.

[7] In Study 3, we also asked participants an additional exploratory question about their political orientation: "When it comes to politics, do you usually think of yourself as liberal, conservative, or something else?" (Choice options: Very liberal; Liberal; Slightly liberal; Moderate/Middle-of-the-road; Slightly conservative; Conservative; Very conservative; Libertarian; Don't know/Not political; Other). Participants' political orientation did not influence the effects of our manipulation, and so we did not include this question in any subsequent studies, and we do not discuss it further.

[8] Before predicting how many points they thought their predicted winner would win by, we also asked participants to indicate how

confident they were that their predicted winning team would win (1 = not at all confident; 7 = extremely confident). This was asked only before participants made their first guess and not before they made their second guess. At the end of the survey, we presented participants with a set of six knowledge questions about NBA basketball. Specifically, we asked them to identify the teams of four different players and to identify which teams had the best and worst records at the time of the study. They were asked to answer these questions without looking up the answers.

[9] Past research on the wisdom of the inner crowd has tended to use more continuous measures of accuracy, such as average absolute (or squared) deviations from the truth (Vul and Pashler 2008; Herzog and Hertwig 2009, 2014b). We preferred (and preregistered) this binary measure for two reasons. First, we were worried that the effects of our manipulation on average absolute (or squared) deviations may be distorted by outliers, and we did not want to make difficult and necessarily arbitrary decisions about how to handle outliers. Second, we were intrinsically interested in the question of how *often* the average guess outperforms the first guess, because the answer to *this* question determines whether it is usually a good idea to rely on first guesses or on average guesses. Nevertheless, we provide the results for a more continuous measure of accuracy in the e-companion.

[10] In each study, we tried to design the survey to make it impossible for participants to enter identical first and second guesses. However, in Studies 1 and 3, there was an error that allowed some participants to do so. We did preregister to drop identical estimates if somehow the software allowed them to enter identical estimates.

[11] As noted earlier, Studies 1 and 5 included an additional third condition. Here we focus our analyses only on the comparison between the Too High/Too Low condition and the Estimate Only condition. We include the results of the third condition in the e-companion.

[12] As mentioned previously, in Study 3, we manipulated whether participants were incentivized for only their second guesses or for the better of their two guesses. Although we found that incentivizing participants for only their second guesses reduced the wisdom of the inner crowd, $b = -0.036$, SE = 0.010, $t(1,259) = -3.57$, $p < 0.001$, and decreased the probability that participants' second guesses would be in the right direction, $b = -0.060$, SE = 0.010, $t(1,259) = -5.82$, $p < 0.001$, there was no interaction between incentive scheme and the Too High/Too Low condition on either of these variables ($ps \geq 0.199$). As shown in Table 2, our results were weakest for Study 4, the study that did not provide any incentives for accuracy. Thus, if anything, it seems that our results are stronger when participants are trying harder to provide correct forecasts.

[13] This will occur so long as the mean of the distribution from which their guesses are drawn is not itself very extreme (with respect to the scale midpoint). If the mean is very extreme, then we would expect people's second guesses to be in a more extreme direction 50% of the time. The R code for these simulations can be found here: https://researchbox.org/45.

[14] This belief was also bolstered by mediation analyses, which suggested that the effects of our manipulation on the two other wisdom-of-the-inner-crowd variables were mediated by this *extreme direction of adjustment* variable. Because mediation analyses are correlational and thus necessarily tentative, we put these analyses in the e-companion.

[15] To understand why this is, consider that participants' first guesses are often on the wrong side of the scale midpoint (between 43% and 47% of the time in Studies 1–4 and 22% of the time in Study 5). All of these wrong-direction first guesses are *too extreme*, as in these cases, participants would benefit from making second guesses that retreat back toward the scale midpoint.

[16] This analysis includes the results from all five studies, but it should arguably exclude the results from Study 5, because what counts as extreme on the points dimension asked about in Study 5 is probably different than what counts as extreme on the percentage dimension asked about in Studies 1–4. When we excluded Study 5 from this analysis, the correlation was virtually identical: $r(38) = -0.48$, $p = 0.001$. Within Study 5 itself, the correlation was large and negative, although only marginally significant because of the small number of items: $r(6) = -0.65$, $p = 0.083$.

[17] Out of habit, we mistakenly preregistered to include fixed effects for prediction question. However, because we did not manipulate the extremity of the true answer within question, it is not possible to conduct our main analysis while including fixed effects for question. We did include these fixed effects in the analyses of the simple effects reported in Table 4 (i.e., in our separate analyses of the questions with moderate versus extreme answers).

[18] It is worth emphasizing just how lucky we were that our first five studies happened to include questions that tended to have moderate answers. If half of the questions had had moderate answers and half had had extreme answers, we probably would have observed a null effect, and erroneously concluded that the Too High/Too Low question had no effect on the wisdom of the inner crowd.

[19] After participants made all their first and second guesses, we asked them the same exploratory questions as we did in the previous studies. At the end of the survey, we also asked them to report how much they know about stocks in general (1 = nothing; 7 = a lot).

[20] In each study, we tried to design the survey to make it impossible for participants to enter identical first and second guesses. However, in Studies 1, 3, and 8, there was an error that allowed some participants to do so. In Studies 1 and 3, we did preregister to drop identical estimates if somehow the software allowed them to enter identical estimates. In Study 8, we neglected to preregister this. We nevertheless excluded these observations from our analyses in Study 8 to be consistent with the prior studies. If we include these participants the results for the extremity and right-direction measures remain unchanged, and the results for the inner crowd wisdom measure become slightly stronger, $b = -0.018$, SE = 0.011, $p = 0.099$.

[21] Because this measure was not preregistered, this method for removing outliers was also not preregistered. However, it is the only outlier removal method that we tried.

[22] We thank an anonymous reviewer for pointing this out.

References

Ariely D, Au WT, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten TS, Zauberman G (2000) The effects of averaging subjective probability estimates between and within judges. *J. Experiment. Psych. Appl.* 6(2):130–147.

Ashton RH (2000) A review and analysis of research on the test-retest reliability of professional judgment. *J. Behav. Decision Making* 13(3):277–294.

Dawes RM (1979) The robust beauty of improper linear models in decision making. *Amer. Psych.* 34(7):571–582.

Galton F (1907) Vox populi. *Nature* 75:450–451.

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psych. Sci.* 20(2):231–237.

Herzog SM, Hertwig R (2014a) Harnessing the wisdom of the inner crowd. *Trends Cognitive Sci.* 18(10):504–506.

Herzog SM, Hertwig R (2014b) Think twice and then: Combining or choosing in dialectic bootstrapping? *J. Experiment. Psych. Learning Memory Cognition* 40(1):218–232.

Hourihan KL, Benjamin AS (2010) Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. *J. Experiment. Psych. Learn. Memory Cognition* 36(4):1068–1074.

Kelly TF, Simmons JP (2016) When does making detailed predictions make predictions worse? *J. Experimental. Psych. General* 145(10):1298–1311.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Sociol. Psych.* 107(2):276–299.

Moore DA, Small DA (2007) Error and bias in comparative judgment: On being both better and worse than we think we are. *J. Personality Sociol. Psych.* 92(6):972–989.

Simmons JP, Nelson LD (2006) Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *J. Experiment. Psych. General* 135(3):409–428.

Simmons JP, Nelson LD (2020) Six biases that are all the same: An introduction to comparative choice theory. Working paper, University of Pennsylvania, Philadelphia.

Steegen S, Dewitte L, Tuerlinckx F, Vanpaemel W (2014) Measuring the crowd within again: A pre-registered replication study. *Frontiers Psych.* 5:786.

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Little Brown, London).

Treynor JL (1987) Market efficiency and the bean jar experiment. *Financial Analysis J.* 43(3):50–53.

van Dolder D, van den Assem MJ (2018) The wisdom of the inner crowd in three large natural experiments. *Natural Human Behav.* 2(1):21–26.

Vul E, Pashler H (2008) Measuring the crowd within: Probabilistic representations within individuals. *Psych. Sci.* 19(7): 645–647.

Winkler RL, Clemen RT (2004) Multiple experts vs multiple methods: Combining correlation assessments. *Decision Analysis* 1(3): 167–176.

Yaniv I (2004) The benefit of additional opinions. *Current Directions Psych. Sci.* 13(2):75–78.