Journal of Experimental Psychology:
Learning, Memory, and Cognition
2014, Vol. 40, No. 1, 218–232

0278-7393/14/$12.00 DOI: 10.1037/a0034054

Think Twice and Then: Combining or Choosing in Dialectical Bootstrapping?

Stefan M. Herzog and Ralph Hertwig
Max Planck Institute for Human Development, Berlin, Germany

Individuals can partly recreate the "wisdom of crowds" within their own minds by combining nonredundant estimates they themselves have generated. Herzog and Hertwig (2009) showed that this accuracy gain could be boosted by urging people to actively think differently when generating a 2nd estimate ("dialectical bootstrapping"). Although the "crowd within" promises accuracy gains, it remains unclear whether and when people spontaneously reap those gains. What makes people combine their estimates rather than trying to identify the better one? This research found that participants were more likely to combine when they were instructed to actively contradict themselves. Furthermore, they were more likely to combine as the size of the disagreement between 1st and 2nd estimate grew. People thus acted as if they were hedging against the risk of making large errors. Finally, when people pursued a strategy other than combination, they were not able to outperform their crowd within.

Keywords: estimation, judgments under uncertainty, dialectical bootstrapping, judgment aggregation, wisdom of crowds

In his classic polemic *Psychologie des Foules* (published in English as *The Crowd*), the French writer Gustave Le Bon (1895/1995) scorned the idea that groups should be allowed to make political decisions. A self-appointed critic of democratic political movements, he likened crowds to "beings belonging to inferior forms of evolution" (Le Bon, 1895/1995, p. 56) and stated that "the crowd is always intellectually inferior to the isolated individual" (p. 53). Crowds are, of course, not always wise. But as we now know, Le Bon thoroughly misjudged the intellectual prowess of groups. Crowds can make decisions superior to those of isolated individuals (Page, 2007)—a phenomenon referred to as the *wisdom of crowds* (Surowiecki, 2004). Even more surprising, an isolated individual can boost his or her performance by simulating what would have appalled Le Bon: a crowd of diverse opinions within his or her own mind (e.g., Herzog & Hertwig, 2009; Vul & Pashler, 2008). However, it seems that we humans share some of Le Bon's sentiments to the extent that we do not necessarily welcome a diversity of opinions within our minds and may thus miss the opportunity to benefit from them. The latter issue is the topic of this article.

The Crowd-Within Effect

Several researchers have demonstrated the benefits of a *crowd-within effect* (Vul & Pashler, 2008) by averaging "quasi-independent" estimates from the same person (Herzog & Hertwig, 2009; Hourihan & Benjamin, 2010; Müller-Trede, 2011; Rauhut & Lorenz, 2011; Stroop, 1932; Vul & Pashler, 2008; Winkler & Clemen, 2004). However, because it is difficult to liberate oneself from the anchor set by one's previous estimate, the errors of n estimates from the *same* individual are likely never as independent as the errors of n estimates from different individuals. Therefore, the averaging gains due to error cancelation are likely to be larger for real crowds than for simulated "crowds within" (Ariely et al., 2000; Herzog & Hertwig, 2009; Rauhut & Lorenz, 2011; Vul & Pashler, 2008; Winkler & Clemen, 2004). However, it is possible to boost the wisdom of the crowd within by making the errors of multiple estimates by the same person less dependent. There are various ways of at least partially releasing a second estimate from the control exercised by the first. Vul and Pashler (2008) showed that introducing a time delay between first and second estimates in response to general knowledge questions increased independence of errors and thus aggregation gains—presumably because forgetting reduced the mnemonic accessibility of the first estimate and the knowledge retrieved at the time to construct it.

Dialectical Bootstrapping: Improving Judgment by Contradicting Oneself

Beyond mere passage of time, another way to foster the wisdom of the crowd within is to explicitly encourage the production of diverse estimates and, by extension, diverse errors within the same person (Herzog & Hertwig, 2009). We have proposed a simple mental tool—*dialectical bootstrapping*—to simulate a *dissonant* crowd within one's mind (Herzog & Hertwig, 2009). Dialectical bootstrapping promises to enhance the quality of quantitative

This article was published Online First September 9, 2013.

Stefan M. Herzog and Ralph Hertwig, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

We thank the Swiss National Science Foundation for Grant 100014_129572/1 to both authors and Laura Wiles and Susannah Goss for editing the article.

Correspondence concerning this article should be addressed to Stefan M. Herzog, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: herzog@mpib-berlin.mpg.de

estimates by exploiting people's capacity to construct dissonant realities in their mind and to then average them. For illustration, take an off-the-wall question such as "how many hairs does a person have on his or her head?" (Santos, 2009). Dialectical bootstrapping assumes that it is possible for a respondent to reduce her error by averaging her initial estimate with her second one. This second, *dialectical* estimate recruits somewhat different assumptions or knowledge, thus rendering the two estimates' errors less dependent. For instance, in the initial estimate, a person may guess the density of hair on a human head and how much area is covered by hair, and then multiply both quantities. When asked to generate a second, dialectical estimate, the person may dig deeper and wonder whether her belief about, say, density, was too low. Based on a revised figure of density, she then generates a second estimate. Averaging both estimates is likely to produce a gain in accuracy relative to the first estimate whenever it is not clear which assumptions are more accurate—the first or the second set.

In order to boost people's ability to recruit different knowledge and assumptions in the process of generating a second, dissonant estimate, Herzog and Hertwig (2009) adopted the *consider-the-opposite* instruction (Lord, Lepper, & Preston, 1984) as one of several possible dialectical techniques. They instructed participants 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 & Hertwig, 2009, p. 234)

Using this secular variant of Oliver Cromwell's exhortation "I beseech you, in the bowels of Christ, think it possible that you may be mistaken" (Carlyle, 1855, p. 448), Herzog and Hertwig (2009) provided an existence proof of the crowd-within effect: By encouraging people to actively think differently, one can nudge them to simulate a virtual crowd within the mind, leading to a boost in accuracy (see also Müller-Trede, 2011). Specifically, the aggregation of first and second estimates yielded greater gains in accuracy relative to a baseline condition in which respondents generated a second estimate without being instructed to "consider the opposite" (see the reliability condition in Herzog and Hertwig; but see also Herzog & Hertwig, 2013; White & Antonakis, 2013). The potential of dialectical bootstrapping should not be reduced to Cromwell's exhortation, that is, the consider-the-opposite instruction; rather, "any elicitation procedure that taps into somewhat nonredundant, yet plausible knowledge is potentially capable of eliciting effective dialectical estimates" (Herzog & Hertwig, 2009, p. 236). The key to dialectical bootstrapping is to have people ponder the same question from different angles. Adopting a different perspective ideally results in estimates whose errors cancel each other out.

Do People Apply Dialectical Bootstrapping Intuitively?

Although research on the crowd-within effect has identified conditions under which combining multiple estimates from a single person works especially well (e.g., increased time delay; Vul & Pashler, 2008), little is known about the extent to which people *spontaneously* apply this error-cancelation tool. In the absence of explicit instructions, will people who are of two minds aggregate their conflicting opinions? Or will they aim to identify the more accurate one? In what follows, we propose two complementary hypotheses as to how people respond to disagreement within their own mind.

Suppose a person first estimates that humans have, on average, 100,000 hairs on their head. Following the consider-the-opposite instruction, she then re-estimates this number to be 150,000. There are at least three ways to now come up with a final estimate. First, she can opt for one of the two estimates, hoping to choose the more accurate of the two. Second, she can determine an intermediate value between 100,000 and 150,000 by somehow combining the two estimates, with this final value representing a (weighted) average. Combining conflicting estimates is thus a kind of diversification strategy, in which a decision maker hedges the risk of choosing the wrong estimate by creating an "error portfolio within one mind." In a special case of such a combination strategy, a person can assign equal weight to both estimates and simply average the two numbers (i.e., 125,000). Third, the person can abandon both estimates and start afresh—rendering a completely new estimate, possibly arriving at a new value located outside the initial range of 100,000 to 150,000. Before we explicate the two hypotheses tested in this article, let us briefly review the only published study to have investigated how people deal with conflicting estimates within one mind (Müller-Trede, 2011).

In Müller-Trede's (2011) investigation, participants applied dialectical bootstrapping to a set of general knowledge questions. They were then presented with their first and their second (dialectical) estimates and asked to give a final estimate. Two findings are particularly relevant for our purposes. First, most people were consistent in how they derived their final estimates. Second, about half the participants tended to combine their first and second estimates, about a quarter tended to choose one of their estimates or to switch between strategies, and another quarter tended to produce final estimates that were outside the range of their initial estimates (see his Figure 2, p. 288). The latter participants presumably "started from scratch," that is, abandoned their first two estimates, retrieved new knowledge, and then rendered a new estimate outside the range defined by their initial estimates.[1]

These results provide first insights into what people appear to do in light of dissonant estimates, but they also leave important questions unanswered. Notably, Müller-Trede (2011) did not employ a reliability (control) condition. Consequently, it is unclear to what extent the modal preference—combining conflicting estimates—hinges on the consider-the-opposite instruction or would also have arisen without it. Because this instruction makes people aware of legitimate alternative opinions on the same issue, it may simultaneously lead them to combine estimates. Another question left open by this first study is what triggers people to combine their opinions or to choose between them. Specifically, although Müller-Trede's results show consistent individual differences in how people deal with their conflicting opinions, it may be that the propensity to combine or to choose is a function of the degree to

[1] Let us emphasize that the "starting from scratch" interpretation is only a working hypothesis at this point. However, it seems clear that participants whose final estimates were outside the range did not combine or choose between their first and second estimates (see also Soll & Mannes, 2011, for further discussion of final estimates outside the range).

which people disagree with themselves. We now formulate two mutually compatible hypotheses as to how people respond to the presence of two dissonant self-generated estimates.

The Diversity-Inclusion Hypothesis

Herzog and Hertwig's (2009) consider-the-opposite instruction encouraged judges to produce two different views on the same issue. In the process of generating these views, people may develop an appreciation for conflicting, but nevertheless reasonable, arguments. Combining the different views into one may thus offer a simple and elegant way of respecting the merits of the different arguments and giving each view a voice. The *diversity-inclusion hypothesis* therefore predicts that the propensity to combine estimates is more pronounced when a person is explicitly instructed to think twice (i.e., in the dialectical condition) than when no such instruction is given (i.e., in the reliability condition).

The Hedging Hypothesis

According to our second hypothesis, people who disagree with themselves "hedge their bets" by combining their conflicting estimates. Such a "combining strategy" can be seen as managing the risk of choosing the less accurate of two estimates (i.e., risk diversification; see Larrick & Soll, 2006). The more people disagree with themselves (i.e., with increasing quantitative difference between the two estimates), the larger the worst-case error should they choose the wrong estimate. Assuming that most people are averse to the risk of making large errors and because combining reduces that risk (e.g., Armstrong, 2001; Hibon & Evgeniou, 2005; Larrick & Soll, 2006), we hypothesize that people become more inclined to combine their estimates, the more they disagree with themselves. Although people's intuitions about averaging do not correspond to normative theories of judgment aggregation (Larrick & Soll, 2006; Soll, 1999), large internal disagreements may prove to be a forceful trigger of aggregation.

Because the dialectical technique aims to produce diverging estimates with ideally uncorrelated errors, the diversity-inclusion and the hedging hypotheses both predict more combining in the context of dialectical bootstrapping than when people simply give another estimate. Unlike the diversity-inclusion hypothesis, however, the hedging hypothesis predicts that the tendency to combine increases with the size of the disagreement (irrespective of whether it stems from the dialectical technique or simply from giving another estimate). Thus, a critical test of the diversity-inclusion hypothesis is whether the dialectical technique still has an influence when the size of the disagreement within a person is controlled.

Diversity Inclusion or Hedging: An Experimental Test

In order to test the diversity-inclusion and the hedging hypotheses, we combined the dialectical bootstrapping paradigm (Herzog & Hertwig, 2009) with the advice-taking paradigm (Bonaccio & Dalal, 2006; see also Müller-Trede, 2011). The latter comprises three steps: (a) determining one's initial opinion, (b) receiving advice, and (c) possibly revising one's initial opinion in light of that advice. Adopting this three-step process, we asked participants, first, to answer a set of general knowledge questions. They were then asked to answer the same questions again under two different conditions. In the *dialectical condition*, they were shown their original answers and instructed to employ the consider-the-opposite strategy. In the *reliability condition*, they received no special instructions (Herzog & Hertwig, 2009). The aim of this condition was to approximate the situation in which participants merely "sample" a second estimate from (roughly) the same subjective probability distribution underlying their first estimate (see Herzog & Hertwig, 2009; Vul & Pashler, 2008). Due to random error, the first and second estimates are likely to vary somewhat, and some aggregation gain is possible. Presenting the first estimates again, as we did in the dialectical condition, would likely have attenuated the independence of the errors and thus the potential for aggregation gains. Therefore, we did not present participants in the reliability condition with their first estimates again (see also Herzog & Hertwig, 2009; Vul & Pashler, 2008).[2] Finally, in the third and final step, participants in all conditions were presented with the questions along with their first and second estimates and were asked to give a final estimate.

In order to test the diversity-inclusion and the hedging hypotheses, we analyzed (a) how often participants chose one of their previous two answers as their final estimate, (b) how often they combined them, (c) how much they deviated from an equal weighting scheme, and (d) how the size of the gap between first and second estimates influenced how participants resolved their conflicting opinions. Furthermore, we investigated whether participants were able to perform more accurately than their own crowd within by identifying the better answer. Finally, we conducted exploratory analyses of the time participants took to come up with their final estimate to shed additional light on their revision strategies.

Beyond testing the diversity-inclusion and the hedging hypotheses, we examined whether applying the consider-the-opposite instruction works repeatedly, that is, whether dialectical bootstrapping can improve estimates that have already been "debiased" (Larrick, 2004). Applying the instruction to their first estimate may leave judges no room to contradict themselves a second time, such that there is no gain to be reaped by combining two dialectical estimates. To test whether "double bootstrapping" is indeed redundant, we introduced the d^2 *condition* (i.e., "double dialectical condition"), in which participants were instructed to think differently (through the consider-the-opposite instruction) when generating both their first and second estimates. Assuming effective debiasing, we predicted that the first estimates in the d^2 condition would be more accurate than the first estimates in the other two conditions. However, no differences were predicted in how participants in the dialectical and d^2 conditions would generate their final estimates.

[2] This setup confounds the dialectical instruction with the presentation of the first estimate in the second step of the study. Although this confound was motivated by the aim of not handicapping the reliability condition with respect to aggregation gains, it may invite alternative interpretations of some of the results. We therefore address this confound experimentally below.

Method

Participants

Two hundred eighty-five students from the University of Basel participated. On average, they were 23.1 years old, interquartile range or IQR [20.0, 24.0]; 79% were female. They received course credit or a flat fee of CHF 10 (about U.S.$9.20 at the time) for participating and could win up to an additional CHF 6 (about U.S.$5.50), depending on the accuracy of their estimates (the payoff scheme is detailed below). On average, they earned an additional CHF 2.56, IQR [2.20, 3.00], about U.S.$2.40. Furthermore, participants were entered in a lottery for a chance to win eight movie vouchers of CHF 20 each (about U.S.$18.50); all participants had the same chance of winning.

Materials and Procedure

Participants responded to 20 general knowledge questions (see Appendix A), each requiring a percentage number as an answer (e.g., "What percent of the world's population aged 15 years or older can read and write?"). We translated all eight items (Vul, 2008) used by Vul and Pashler (2008) into German and created 12 more items based on facts reported in *The World Factbook* (Central Intelligence Agency, 2008). The study was run on computers.

Participants were randomly assigned to one of three conditions. In the *reliability* ($n = 95$) and *dialectical* ($n = 95$) conditions, they first generated their best estimates (without knowing that they would be asked the same questions again). In the *d^2 condition* (i.e., "double dialectical condition"; $n = 94$), participants were instructed to produce their first estimate by applying the consider-the-opposite strategy (also without knowing that they would be asked the same questions again). Specifically, they were instructed as follows (translated from German):

> First, please make an estimate. But don't enter it into to the computer yet. Second, think about a few reasons why your initial estimate could be wrong. Third, now make your best estimate and enter it into the computer.

These instructions were meant to "debias" participants' first estimates as envisioned by debiasing research (Larrick, 2004) through application of a consider-the-opposite strategy (see Lord et al., 1984).

In the second phase of the experiment, participants in the reliability condition made a second estimate without the consider-the-opposite instruction and without their first estimates being displayed. Specifically, they were instructed as follows (translated from German; see also Herzog & Hertwig, 2009):

> Imagine you are starting this study now and that you are making your estimates for the first time. For each question, please give your best estimate.

Participants in the dialectical and the d^2 conditions were asked to give a dialectical estimate using the consider-the-opposite strategy and with their first estimate being displayed in front of them. Specifically, they were instructed as follows (translated from German; see Herzog & Hertwig, 2009, p. 234):

> 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.

Finally, participants in all three conditions were presented for a third and final time with each question and their two previous answers and were asked to give a final estimate: "You now have the opportunity to consider your previous two estimates while making a third and final estimate for each question" (translated from German). The order of the items was randomized for each participant but kept constant across the three blocks of estimates within each participant.

Participants received CHF 0.10 (about U.S.$0.09) for each answer with a lower absolute error than the median absolute error from a pool of answers collected previously. This payment scheme was announced at the beginning of the study. Prior to their second estimates, participants were told that the better of their first two estimates for each question would determine their payoff in the second phase. Answers in the third phase, like those in the first phase, were incentivized in isolation. This payment scheme aimed at encouraging participants in the dialectical and d^2 conditions to give dissimilar second estimates. To render all conditions comparable, we used the same incentive scheme in the reliability condition (see Herzog & Hertwig, 2009). Participants received no performance feedback throughout.

Statistical Analysis

Our accuracy measure is mean absolute deviation (*MAD*). We chose this measure of the quality of estimates because it is widely used in the advice-taking literature (e.g., Soll & Mannes, 2011; Yaniv & Kleinberger, 2000) and because it does not favor averaging over choosing to as great an extent as, for example, mean squared deviation (*MSD*) does (by heavily punishing large errors; see, e.g., Soll & Larrick, 2009, p. 784). All averages of two estimates were rounded, so that any superiority of the averages did not result from them being more fine-grained than the raw estimates.

With the one exception stated above (accuracy of the first estimates), we did not predict any differences between the dialectical and the d^2 conditions. We therefore pooled the two dialectical conditions and report contrasts between them, on the one hand, and the reliability condition, on the other, unless another analysis was more appropriate.

We used a Bayesian parameter estimation approach with vague priors to analyze the data (e.g., Kruschke, 2011a, 2011b). To this end, one (a) selects an appropriate descriptive statistical model of the data (e.g., normal distribution), (b) postulates prior distributions of the parameters (e.g., mean and standard deviation of a normal distribution) acceptable to a skeptical audience, and (c) updates those prior distributions in the light of the data, using Bayes theorem. Vague priors are quickly overruled by data and thus do not have a substantial influence on inference. The resulting posterior distributions of the parameters represent what one should believe about the parameters after having seen the data. We report the mode of a posterior and the 95% highest posterior density interval (HDI); the HDI indicates the parameter range "for which all values inside the interval have higher credibility than values outside the interval, and the interval contains 95% of the distribution" (Kruschke, 2011a, p. 302). In Appendix B, we report the methods in more detail (for information on the strengths of a

Bayesian approach to statistics see, for example, Dienes, 2011; Kruschke, 2010, 2011a, 2011b; Wagenmakers, 2007).

Results and Discussion

Table 1 gives an overview of the main results for the three conditions separately, whereas in the following we report linear contrast analyses between the *two* dialectical conditions and the reliability condition as specified above.[3]

Preliminary Analyses

First, measured in MADs, participants' first, debiased estimates in the d^2 condition were more accurate than the first estimates in the reliability and dialectical conditions by 1.56 percentage points, 95% HDI [0.64, 2.64], $d = 0.42$ (see also row MAD_1 in Table 1). Thus, the debiasing of first estimates in the d^2 condition was successful.

Second, participants in the two dialectical conditions produced first and second estimates that were further apart than those of participants in the reliability condition. In terms of a participant's median absolute difference between estimates across items, the answers differed by 10.44 percentage points, HDI [9.69, 11.24], in the dialectical conditions relative to 6.66, HDI [6.04, 7.29], in the reliability condition (Cohen's $d = 0.91$; see also row Median AD_{12} in Table 1).

Third, the consider-the-opposite instruction used in the dialectical and d^2 conditions boosted the independence of errors, relative to the reliability condition. We used the bracketing rate (Larrick & Soll, 2006) as a measure of error independence (i.e., the proportion of questions for which the correct answer was between the first and second estimate and part of the error was thus canceled when the two were averaged; higher values indicate less correlated errors). The dialectical conditions showed higher bracketing rates than the reliability condition did (see also row Bracketing rate in Table 1): 22%, HDI [0.20, 0.23], versus 14%, HDI [0.12, 0.16]; Cohen's $d = 0.94$.

Fourth, dialectical bootstrapping increased accuracy somewhat (see Table 1)—also when the first estimates were already generated under the consider-the-opposite instruction (i.e., d^2 condition). In the reliability condition, averaging first and second estimates did not reduce error relative to first estimates: The MAD of the average of the first and second estimates was, on average, 0.14%, HDI [–1. 58, 1.72], $d = 0.02$, larger than the MAD of the first estimates. In contrast, averaging first and second estimates in the dialectical and d^2 conditions reduced the MAD of the average, relative to the MAD of the first estimate, by 1.89%, HDI [0.47, 3.14], $d = 0.22$.[4] The posterior probabilities that those accuracy gains are positive were 47% and 99.6% for the reliability and the two dialectical conditions, respectively. The relative error reduction in the dialectical conditions was 1.96 percentage points higher than in the reliability condition, HDI [–0.16, 4.05], $d = 0.23$; the posterior probability that this difference is positive was 96%.

Modeling the Final Estimate

We modeled a participant's third estimate (e_3) for each question as a weighted average of the first (e_1) and second (e_2) estimates as follows:

$$e_3 = w_1 \times e_1 + (1 - w_1) \times e_2,$$

where w_1 is the weight a participant places on her first estimate. Rearranging terms gives $w_1 = (e_2 - e_3)/(e_2 - e_1)$ for the weight placed on the first estimate (see Bonaccio & Dalal, 2006, p. 141). When participants offered the same quantity for the first and second estimate, then w_1 is undefined; this happened in 15%, 1%, and 3% of cases in the reliability, dialectical, and d^2 conditions, respectively. Participants' third and final estimate was situated outside the range of the first two estimates in 20%, 19%, and 20% of trials in the reliability, dialectical, and d^2 conditions, respectively; the latter three quantities are lower than the 30% reported in Müller-Trede (2011, p. 288).

Following Müller-Trede (2011, p. 288), we calculated for each item and each participant a quantitative measure of how much the final estimate deviates from an equal-weight strategy as follows:

$$\Delta = \left| w_1 - 0.5 \right|.$$

Replicating his result (Müller-Trede, 2011, p. 288), we found consistent individual differences in how much participants deviated from such an equal-weight strategy. As a measure of reliability, we calculated Cronbach's α and found it to be 0.63 in the reliability condition, 0.79 in the dialectical condition, and 0.83 in the d^2 condition. The αs in the latter two conditions were thus close to the 0.85 reported by Müller-Trede (2011). However, participants in our reliability condition were markedly less consistent (i.e., 0.63; Müller-Trede, 2011, had no reliability condition).

In the light of these basic results, let us briefly reiterate our two hypotheses. The diversity-inclusion hypothesis predicts that when people explicitly take different perspectives on the same question (induced by the consider-the-opposite instruction), they are more inclined to combine the two resulting estimates than are people in the reliability condition. The hedging hypothesis predicts that people are more inclined to combine their two estimates the further apart those estimates are. We now test both predictions.

Testing the Diversity-Inclusion Hypothesis

For each participant, we calculated the median Δ across items (see Müller-Trede, 2011, p. 288) and then classified the participant to different *revision strategies* as follows. Participants who combined their estimates on most trials (i.e., median $\Delta < 0.5$, that is, in at least 50% of the trials Δ was smaller than 0.5) were classified as users of the "combining strategy." Among those users, we additionally categorized those who put roughly equal weight on both their estimates in the majority of trials as users of the "averaging strategy." Specifically, following Soll and Larrick (2009), we treated weights between 40% and 60% as tantamount to averaging (i.e., median $\Delta < 0.1$). Participants with final answers outside the range of their first and second estimates on most trials were classified as users of the "starting-from-scratch strategy" (i.e., median $\Delta > 0.5$). We suspect that most such estimates stem from participants abandoning their previous two estimates, retriev-

[3] Note that because of the Markov Chain Monte Carlo (MCMC) implementation of the Bayesian statistics (see Appendix B for details), the mode of a posterior distribution of a linear contrast need not coincide with a linear contrast value that is calculated using the modes of the posterior distributions of the individual conditions (as presented in Table 1).

[4] The percentage improvement was calculated on the participant level as $(MAD_1 – MAD_{avg12})/MAD_1$ and then summarized across participants (analogously to the symmetrical case in Larrick & Soll, 2006; Soll & Larrick, 2009, where averaging is compared with choosing randomly between all first or all second estimates).

Table 1
Descriptive and Accuracy Measures for the Reliability, Dialectical, and Double Dialectical (d^2) Conditions

Measure	Reliability condition				Dialectical condition				d^2 condition				
	M	*SD*	95% HDI	d_0	*M*	*SD*	95% HDI	d_0	*M*	*SD*	95% HDI	d_0	log10 (*v*)
Median AD_{12}	6.66	2.81	[6.04, 7.29]		11.15	4.94	[10.05, 12.31]		9.77	4.42	[8.76, 10.75]		0.98
MAD_1	18.70	4.03	[17.84, 19.56]		18.51	3.83	[17.69, 19.32]		16.98	3.79	[16.16, 17.78]		1.54
MAD_2	19.88	4.04	[18.97, 20.72]		19.56	4.57	[18.56, 20.64]		17.97	3.84	[17.11, 18.81]		1.01
MAD_2–MAD_1	1.12	2.59	[0.54, 1.73]	0.44	0.86	3.13	[0.16, 1.69]	0.28	0.90	2.24	[0.38, 1.41]	0.40	0.79
MAD_{avg12}	18.64	3.82	[17.84, 19.46]		18.16	3.90	[17.28, 18.96]		16.64	3.69	[15.82, 17.40]		1.43
% MAD reduction averaging													
Relative to 1st estimate	−0.14	7.44	[−1.72, 1.58]	−0.02	1.86	9.22	[−0.25, 3.86]	0.19	1.83	7.40	[0.22, 3.52]	0.26	0.95
Relative to 2nd estimate	5.67	7.53	[4.11, 7.35]	0.76	7.14	9.29	[5.11, 9.22]	0.75	7.02	8.94	[4.98, 8.92]	0.78	1.15
Bracketing rate	0.14	0.06	[0.12, 0.16]		0.22	0.09	[0.20, 0.25]		0.21	0.10	[0.18, 0.24]		
Accuracy ratio	1.12	0.08	[1.10, 1.15]		1.11	0.08	[1.09, 1.13]		1.09	0.07	[1.07, 1.11]		0.36
MAD_3	18.93	4.16	[18.01, 19.81]		18.34	3.81	[17.51, 19.17]		17.19	3.80	[16.37, 18.05]		1.24
Median Δ	0.43	0.16	[0.40, 0.47]		0.36	0.19	[0.32, 0.41]		0.37	0.19	[0.33, 0.42]		0.76

Note. Participants answered 20 questions on a percentage point scale, and the first five measures in Table 1 are expressed as percentage points (calculated on the participant level). Median AD_{12} indicates the median absolute distance between first and second estimates. MAD_1, MAD_2, and MAD_3 indicate the mean absolute distance (MAD) between estimates and the correct answers for the first, second, and final estimates, respectively. MAD_2–MAD_1 indicates the increase in MAD when comparing second to first estimates. MAD_{avg12} indicates the MAD of the average of first and second estimates for a question. % MAD reduction averaging indicates the proportional reduction in MAD when comparing the MAD of the average of first and second estimates to the MAD of the first or second estimates, respectively. The bracketing rate is the proportion of items for which the first and the second estimates have an error of different sign, that is, where one estimate overestimates and the other underestimates the true value; the bracketing rate is a measure of the independence of the errors of the first and second estimates (see Larrick & Soll, 2006). The accuracy ratio is calculated as $\max(MAD_1, MAD_2)/\min(MAD_1, MAD_2)$ and indicates the MAD ratio of worse to better set of estimates (see Soll & Larrick, 2009). If the ratio is 1, both sets of estimates are equally accurate ($MAD_1 = MAD_2$); if the ratio is, for example, 1.1, then the worse set's MAD is 10% higher than that of the better set. See the main text for the interpretation of the bracketing rate and the accuracy ratio within the probability, accuracy, redundancy (PAR) model (Soll & Larrick, 2009). For each of these measures, which are all calculated on the participant level first, we calculated statistics for each condition separately: the mean (*M*), the standard deviation (*SD*), the 95% highest density interval (95% HDI) of the mean and the one-sample Cohen's d_0 effect size (compared against zero; reported only where appropriate). Note that because of the Markov Chain Monte Carlo (MCMC) implementation of the Bayesian statistics, the mode of a posterior distribution of a linear contrast (e.g., the two dialectical conditions vs. the reliability condition, as reported in the main text) need not coincide with a linear contrast value that is calculated using the modes of the posterior distributions of the individual conditions presented here. Dash indicates that there is no *v* parameter in the model used for this variable. All distributions were treated as continuous (except for bracketing rates, which were modeled as proportions), and *v* is the normality parameter of a *t* distribution (lower values indicate kurtosis, higher values normality). See Appendix B for more details of statistical procedures. 95% HDI = 95% highest posterior density interval; max = maximum; min = minimum; avg = average.

ing new knowledge, and rendering a novel estimate (and their relatively long response times, see our analyses below, are consistent with this interpretation; see also footnote 1). Finally, we classified participants whose final estimate equaled one of their first two estimates on most trials as users of the "choosing strategy" (i.e., at least 50% of the Δs equaled 0.5, which implies choosing one of the first two estimates). This leaves us with participants who displayed a median Δ of exactly 0.5 but did not reproduce one of their first two estimates on most trials: These participants combined, chose, and went outside the range on various trials and by coincidence ended up with a median Δ of 0.5. We classified these participants as users of the "eclectic strategy."

Figure 1 shows the results of this classification. Several findings are noteworthy: First, the majority of participants in the two dialectical conditions were classified as using the combining strategy (62%), HDI [.55, .68], relative to only 41%, HDI [.31, .50], in the reliability condition—a difference of 21 percentage points. Second, only a few participants strictly averaged their first two estimates. Third, about the same proportions of participants in the reliability condition and the dialectical conditions were classified as choosers: 20%, HDI [.13, .28], and 15%, HDI [.10, .20], respectively. Fourth, there were more users of the eclectic strategy in the reliability condition (26%), HDI [.18, .35], than in the dialectical conditions (11%), HDI [.07, .15], consistent with the lower consistency of participants in the reliability condition (see Cronbach's α analysis). Finally, only a few (roughly 12%) participants consistently started from scratch, as opposed to the 28% reported by Müller-Trede (2011, p. 288).

Our finding of more combining in the dialectical and d^2 conditions relative to the reliability condition is consistent with the notion that people are more likely to recruit the combining strategy when instructed to adopt different views. Relatedly, we also found, on average, lower median Δs (i.e., more equal weights on both estimates) in the dialectical conditions ($M = 0.37$), HDI [0.34, 0.40], than in the reliability condition ($M = 0.43$), HDI [0.40, 0.47]; Cohen's $d = 0.37$. Yet in the former two conditions, first and second estimates were also further apart than in the reliability condition (see analysis above). Consequently, we cannot yet rule out the possibility that the higher proclivity to combine is due to more hedging against large errors (rather than to an appreciation of different views, as conjectured by the diversity-inclusion hypothesis).

Testing the Hedging Hypothesis

Do people tend to put more equal weight on both estimates (i.e., lower Δs) when they are further apart, as suggested by the hedging hypothesis? For the following regression analyses, we operationalized the magnitude of the within-person disagreement in terms of the natural logarithm of the absolute distance between the first and

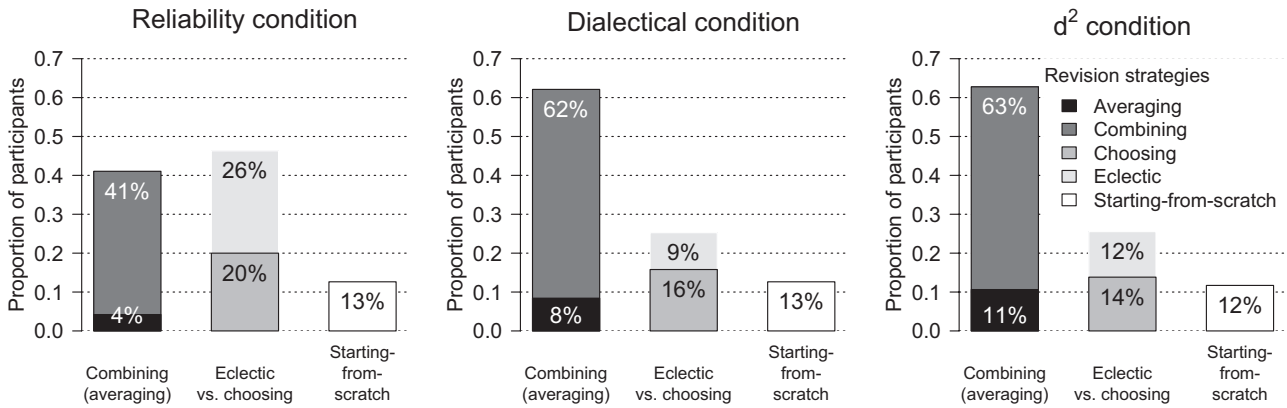

Figure 1. Classification of participants to five revision strategies in the reliability, dialectical, and d^2 conditions. In all three panels, the bars to the left show the proportion of participants who mostly combined their estimates ("combining"), with the lower, black part representing the subset of combiners who mostly averaged their estimates ("averaging"). The middle bars show the stacked proportions of participants who mostly chose one estimate ("choosing": lower bar in grey) or who switched between strategies ("eclectic": upper bar in light grey). The bar to the right shows the proportion of participants who mostly gave final estimates outside the range of their first two estimates ("starting-from-scratch" in white; see main text for a description of the classification procedure). d^2 = double dialectical condition.

second estimates for each item (*logAD12*). We estimated a Bayesian mixed-effects linear model (Baayen, Davidson, & Bates, 2008; Hadfield, 2010) that predicted Δ (dependent variable) simultaneously from (a) an intercept, (b) two indicator variables for the dialectical and the d^2 conditions, respectively, (c) logAD12 (i.e., the size of the disagreement), and (d) the two interaction terms obtained when crossing the two indicator variables with logAD12. In addition to those fixed effects, participant effects for the intercept, the slope of logAD12, and their covariance were estimated.[5] In all regression analyses, we excluded seven extreme trials (0.1%) with a Δ value greater than 10.

Table 2 shows the results. Consistent with the hedging hypothesis, the more strongly participants disagreed with themselves (i.e., the larger logAD12), the more equal weights were put on both estimates (i.e., lower Δ, as indicated by the negative slope for logAD12). In addition, participants in the dialectical and d^2 conditions placed more equal weight on both estimates, relative to participants in the reliability condition (as indicated by the negative estimates for the two indicator variables). Because the regression model simultaneously controls for the magnitude of the disagreement (i.e., logAD12), this finding implies that the consider-the-opposite instruction prompts more equal weights on both estimates *irrespective* of the magnitude of the disagreement. From this follows that the more prevalent use of the combining strategy in the dialectical and d^2 conditions is not solely due to the larger disagreements in those conditions.

In conclusion, our analysis supports both of our hypotheses: Participants put more equal weight on both of their opinions if they were prompted by a dialectical technique (diversity-inclusion hypothesis) and the more they disagreed with themselves (hedging hypothesis).

Table 2
Predicting Deviation From Equal Weighting (Δ) Based on Condition and Size of the Disagreement Within Participants

Parameter	Coefficient	95% HDI
Fixed effects		
Intercept	1.21	[1.03, 1.39]
Dialectical	−0.47	[−0.71, −0.20]
d^2	−0.54	[−0.78, −0.27]
logAD12	−0.32	[−0.38, −0.26]
Dialectical × logAD12	0.18	[0.08, 0.27]
d^2 × logAD12	0.20	[0.11, 0.30]
Participant random effects: (Co)variances		
Intercept	0.60	[0.49, 0.75]
Intercept–logAD12	−0.20	[−0.25, −0.16]
logAD12	0.07	[0.05, 0.09]

Note. Bayesian mixed-effects linear model (Baayen et al., 2008; Hadfield, 2010). 95% HDI indicates the 95% highest posterior density interval. logAD12 = the natural logarithm of the absolute distance between the first and second estimates for each item; d^2 = double dialectical condition.

What Are the Effects of Displaying the Original Estimate?

Participants in the two dialectical conditions were shown their original estimate while they generated their second, dialectal estimate. As explained above, the same did not hold in the reliability condition. It is not inconceivable that the availability of the first estimate may have influenced the way the second estimate was

[5] Preliminary modeling using maximum likelihood estimation showed that, with regard to fixed effects, the inclusion of the interaction terms improved model fit as indicated by the Bayesian information criterion (BIC). Furthermore, regarding participant effects, permitting for variation in the slope for logAD12 improved model fit, whereas condition indicator variables did not. Including a random intercept or a logAD12 slope for items did improve model fit, but the estimated (co-)variances were very small. Because the estimates of fixed effects and participant effects were barely affected by the inclusion or exclusion of item effects, we decided not to estimate them in the Bayesian model.

produced in the two dialectical conditions. Specifically, the larger differences between first and second estimates, the higher bracketing rates, and the higher prevalence of combiners in the dialectical conditions may have been caused by the renewed presentation of the first estimates and have nothing to do with the dialectical instruction per se.

To investigate this possibility, we conducted a control experiment with both (a) the same reliability condition as implemented in the main experiment ($n = 50$) and (b) a "reliability-plus" condition ($n = 50$), which was identical to the reliability condition with one exception: As in the dialectical conditions, each person's first estimates were displayed while he or she generated second estimates. The only other differences from the reliability condition in the main study concerned the participant pool and compensation. Participants were recruited at the Max Planck Institute for Human Development in Berlin. They received a flat fee of €10 (about U.S.$13.3 at the time) and, in addition, could win up to €6 (about U.S.$8) for accurate estimates. On average, participants earned an additional €2.69, IQR [2.30, 3.10], about U.S.$3.60. No lottery was conducted.

If the high prevalence of combining in the dialectical conditions is indeed due to the first estimates being displayed, the results in the reliability-plus condition should mirror those of the dialectical conditions; otherwise, the results in the reliability-plus and reliability conditions should be similar. We found three main results. First, the distance between participants' first and second estimates was similar in the two reliability conditions: In terms of a participant's median absolute difference between the respective first and second estimates across items, the answers differed by 6.44 percentage points ($SD = 2.89$), HDI [5.56, 7.44], in the reliability condition, relative to 5.79 ($SD = 2.80$), HDI [4.94, 6.82], in the reliability-plus condition, HDI of the difference [–1.89, 0.69]; Cohen's $d = -0.20$. Second, the errors between first and second estimates were similar in both conditions. The bracketing rates were 17% ($SD = 6\%$), HDI [0.14, 0.20], in the reliability condition and 14% ($SD = 5\%$), HDI [0.11, 0.17], in the reliability-plus condition, HDI of the difference [–0.07, 0.01]; Cohen's $d = -0.43$. Third, when participants were classified to revision strategies, 46%, HDI [.32, .58], of those in the reliability condition were classified as combiners, relative to only 31%, HDI [.18, .43], in the reliability-plus condition (and 41% in the original reliability condition in the main experiment; see Figure 1).

In sum, the results from the reliability-plus condition did not mirror those observed in the dialectical conditions (see Figure 1). If anything, we found less combining in the reliability-plus condition than in the two reliability conditions in which participants' first estimates were not displayed. The differences observed between the two dialectical conditions and the reliability condition (see Figure 1) thus do not appear to be attributable to the presentation of participants' first estimates in the dialectical conditions.

To Combine or Not to Combine: Can People Outperform the Crowd Within?

We have shown that most participants tended to combine their conflicting estimates when asked for a final answer. Sometimes, however, participants chose either their first or second estimate—or settled on an answer outside the range of their two initial estimates. But did participants forgo combining under the right circumstances? In the following, we report two sets of analyses addressing whether participants' revision behavior was well adapted to the crowd-within context. First, we present an analysis based on the probability, accuracy, redundancy (PAR) model (Soll & Larrick, 2009), which focuses on two strategies for revising quantitative estimates: *choosing* versus *averaging*. Second, we investigate whether or not participants outperformed the crowd within. That is, for those cases in which participants did *not* combine their estimates, we analyze whether or not their final answers were more accurate than a simple combination of their first and second estimates.

PAR model analysis. The PAR model (Soll & Larrick, 2009) was developed to evaluate the revision of quantitative estimates in the advice-taking context; we adapted it for the crowd-within context. The model distinguishes two basic, prototypical strategies for revising estimates across a set of questions: consistently *choosing* the (presumably) better set of estimates (i.e., either always choosing one's first estimate or always choosing one's second estimate) versus consistently *averaging* one's estimates (i.e., averaging one's first and second estimate for each question using equal weights).

According to the model, three conditions must hold for *choosing* to be more accurate than *averaging*: (a) there is a substantial *probability*, p, of selecting the better of two sets of estimates, (b) one set of estimates is clearly better than the other (to capture this difference in accuracy, an *accuracy ratio*, A, is defined as the ratio of the MADs of the two sets of estimates, higher over lower), and (c) the errors in the two sets of estimates must be relatively similar (i.e., low *bracketing rate*, Br).

The PAR model identifies combinations of A and Br for which choosing and averaging are equally accurate. Figure 2 plots iso-accuracy curves for different values of p as a function of A and Br. Averaging outperforms choosing above a given curve and underperforms choosing below it. Several insights can be derived from Figure 2 (see Soll & Larrick, 2009). First, there are situations in which one set of estimates is less accurate than the other (say, $A = 1.3$), but averaging still outperforms choosing even with perfect identification ($p = 1.0$). This is the case when Br is about 35% or higher (see Figure 2). Thus, low error correlation, as indicated by a high bracketing rate, can compensate for sizeable differences in accuracy even when the better estimates can be identified with certainty. Second, with more realistic values of p (i.e., $p < 1$), averaging outperforms choosing for an increasingly larger region of the parameter space. In fact, when p becomes 50%, the iso-accuracy curve coincides with the x-axis. This implies that averaging is always more accurate than choosing (irrespective of p and A) unless the bracketing rate is exactly zero, in which case averaging and choosing are equally accurate (see Larrick & Soll, 2006).

Figure 2 plots the combination of Br and A for each condition (see also Table 1). These combinations indicate the probability, p, with which the typical judge needs to be able to identify the more accurate set of estimates to outperform averaging. Using Figure 2, we can now interpret the values observed in our study. We begin with the bracketing rate. The consider-the-opposite instruction in the dialectical and d^2 conditions produced a higher Br than the reliability condition did: 22%, 21%, and 14%, respectively. The average As were 1.11 in the dialectical, 1.09 in the d^2, and 1.12 in the reliability conditions. In the reliability condition, the inferred p

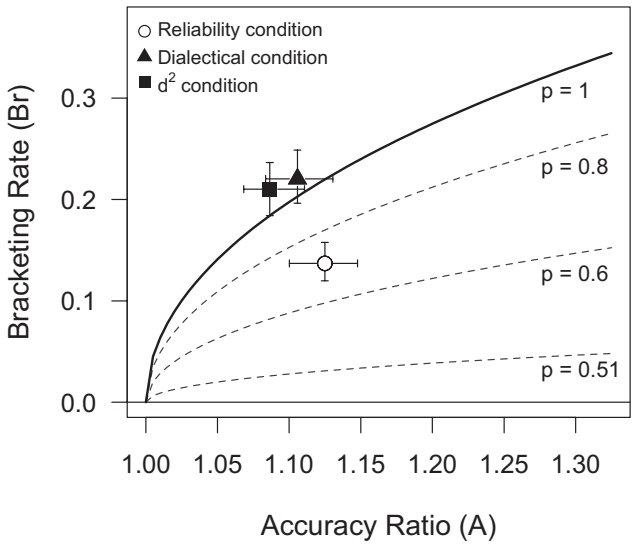

Figure 2. The figure plots accuracy ratios (*A*) and bracketing rates (*Br*) for the reliability, dialectical, and d^2 conditions with the corresponding 95% highest density intervals. The figure also shows iso-accuracy curves for different values of *p* (the probability of selecting the more accurate set of estimates, that is, the first or second set). For a given *p*, consistently *averaging* is more accurate for combinations of *A* and *Br above* the respective curve and consistently *choosing* the more accurate set of estimates is more accurate for combinations *below* the respective curve. d^2 = double dialectical condition. Adapted from Figure 3 in "Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions," by J. B. Soll and R. P. Larrick, 2009, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* p. 785. Copyright 2009 by American Psychological Association.

is located between the iso-accuracy curves for *p*s of .6 and .8. In the two dialectical conditions, the respective values are located above the iso-accuracy curve for a *p* of 1. In other words, in the two dialectical conditions, the typical participant is not expected to outperform averaging by choosing the better set of estimates (either all her first or all her dialectical estimates).

In sum, we took advantage of the PAR model (Soll & Larrick, 2009) to find out whether dialectical bootstrapping creates conditions under which choosing is more effective than averaging, or vice versa. Our analysis revealed that the consider-the-opposite instruction employed in the dialectical and the d^2 conditions boosted error independence, relative to the reliability condition. It did so without increasing the errors of the estimates (i.e., similar accuracy ratios). Consequently, participants in the dialectical and d^2 conditions would need to be very skilled in identifying the more accurate set of estimates for averaging to be outperformed by choosing; they are thus well advised to combine their estimates rather than take the risk of choosing the better set of estimates.

Can people outperform the crowd within? The previous PAR model analysis compared the performance of two hypothetical, "pure" strategies: (a) consistently choosing the presumably better set of estimates (i.e., either all one's first estimates or all one's second estimates) versus (b) consistently averaging one's first and second estimates. As our classification analysis showed, however, not all of our participants followed pure revision strategies (see Figure 1). Therefore, in a second set of analyses, we investigated whether or not participants outperformed the crowd within in those cases in which they did not combine their estimates. We thus focused on those trials in which participants either chose one of their initial two estimates as their final estimate or went outside the range defined by those two estimates. We then compared their final estimates' error (i.e., mean absolute error, MAD) to the error they would have made had they simply averaged instead. Before turning to the results, let us emphasize two things. First, similar results emerged when we analyzed all trials, including those in which participants combined their estimates. Second, combining both estimates using unequal relative to equal weights (averaging) generally leads to similar accuracy gains because of the "flat maximum effect" (whenever the weights are not extreme; Soll & Larrick, 2009); thus, using averaging (i.e., equal weights) is a representative benchmark for this analysis.

Pooled across conditions, only 42%, HDI [0.36, 0.48], of participants beat their crowd within, that is, achieved a lower error by choosing or starting from scratch rather than averaging. Looking at the conditions separately, we found that participants roughly matched the performance of their crowd within in the reliability condition and the dialectical condition (47%, HDI [0.37, 0.57] and 44%, HDI [0.33, 0.54], respectively); in the d^2 condition, however, participants performed worse than their crowd within, 36%, HDI [0.27, 0.46]. Our results are thus broadly consistent with Müller-Trede's (2011) finding. His participants would have performed better had they consistently averaged their first and dialectical estimates.

How Long Does the Third and Final Estimate Take?

Response times can shed additional light on the strategy used to produce the third and final estimate. Earlier, we suggested that large numerical differences between the first and second estimates represent a conflict that participants can reconcile by, for instance, combining or choosing. As resolving conflicts commonly takes time, this process can be expected to take longer, the more strongly people disagree with themselves. Moreover, response times can illuminate our interpretation of outside-the-range estimates as mainly originating from the starting-from-scratch strategy (see also footnote 1): When asked to come up with a third estimate, do people indeed abandon their previous two estimates and judge anew based on newly retrieved knowledge, or do they just ratchet up one of their previous estimates? The starting-from-scratch interpretation suggests that outside-the-range estimates take longer than do final estimates that simply use the previous two estimates as input; if people merely ratchet up one of their previous two estimates, no such difference would necessarily be expected.

In the third and final phase of the experiment, we recorded the time that elapsed between presentation of each item and the participant's entry of her final estimate. We ran two sets of Bayesian mixed-effects linear regression models (Baayen et al., 2008; Hadfield, 2010). In the first model, we predicted the logarithm of the response time (dependent variable) from (a) an intercept, (b) two indicator variables for the dialectical and the d^2 condition, respectively, (c) the size of the disagreement (i.e., logAD12), (d) how much participants deviated from an equal-weight strategy (i.e., Δ), and (e) the interaction between logAD12 and Δ; random intercepts for both participants and items were

estimated.[6] As Table 3 shows, we observed that the more strongly people disagreed with themselves and the more they deviated from an equal-weight strategy, the longer it took them to produce a final estimate (as indicated by the positive slopes for logAD12 and Δ). Second, this slowing down was amplified when both the disagreement and the deviation from an equal-weight strategy were large (as indicated by the positive interaction effect for logAD12 and Δ). Third, the three conditions did not reliably differ in response times when we controlled for the magnitude of the disagreement (logAD12), Δ, and their interaction (see the slopes for the indicator variables for the dialectical and d^2 conditions).

In the previous analysis, we treated Δ as a graded quantitative measure. Next, we categorized each trial for each participant into one of the following four categories, corresponding to qualitatively different ways of resolving the conflict in the crowd within: (a) the final estimate was outside the range of the first and second estimate (starting-from-scratch), (b) the participant chose one of the estimates (choosing), (c) the participant combined the two estimates, with more weight being placed on one of them (i.e., $0 < w_1 < 0.4$ or $0.6 < w_1 < 1$; combining), and (d) the participant averaged the two estimates (i.e., $0.4 < w_1 < 0.6$). Then, separately for each of the three conditions, we ran a mixed-effects linear regression model and predicted the logarithm of the response time (dependent variable) by (a) an intercept, (b) three indicator variables for the four categories described above (setting averaging as the reference category), and (c) the size of the disagreement (i.e., logAD12); random intercepts were estimated for both participants and items.[7]

Figure 3 shows the estimated differences in response time for the different strategies relative to averaging. Three results are noteworthy. First, across all conditions, the production of a third estimate outside the range of the first and second estimates took longer than any other third estimate. Rather than ratcheting up an original estimate, people appear to rethink the issue (i.e., to start from scratch). Second, there is no clear evidence that computationally simpler strategies (i.e., choosing or averaging) take less time than does attributing different weights to both estimates (i.e., combining). And finally, there were no clear differences between combining and choosing or between the three conditions.

Table 3
Predicting Log-Response Speed Based on Condition, Size of the Disagreement Within Participants, and Deviation From Equal Weighting (Δ)

Parameter	Coefficient	95% HDI
Fixed effects		
Intercept	1.99	[1.87, 2.10]
Dialectical	−0.10	[−0.21, 0.02]
d^2	−0.06	[−0.18, 0.05]
logAD12	0.10	[0.08, 0.13]
Δ	0.07	[0.04, 0.11]
logAD12 × Δ	0.05	[0.03, 0.08]
Participant random effects: Variances		
Intercept participants	0.15	[0.13, 0.18]
Intercept items	0.02	[0.01, 0.04]

Note. Bayesian mixed-effects linear model (Baayen et al., 2008; Hadfield, 2010). 95% HDI indicates the 95% highest posterior density interval. logAD12 = the natural logarithm of the absolute distance between the first and second estimates for each item; d^2 = double dialectical condition.

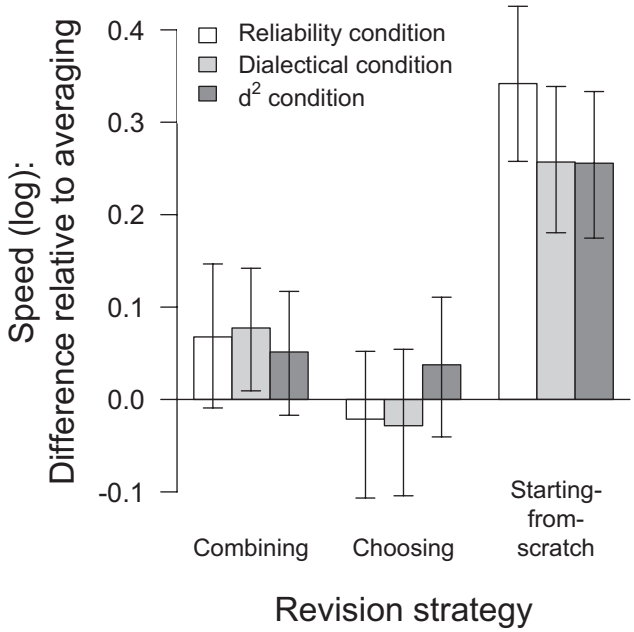

Figure 3. Response times for the combining, choosing, and starting-from-scratch trials, separately for the three conditions. Bars represent the regression estimates for the difference between the revision strategies relative to averaging (i.e., indicator variables), with control for the magnitude of the disagreement within participants (i.e., logAD12); the error bars show 95% highest density intervals (HDI). For each condition, a separate Bayesian mixed-effects linear model was estimated, predicting the log of the response speed in seconds (see main text for details). d^2 = double dialectical condition.

In sum, our exploratory response time analyses suggest two main insights. First, the more strongly people disagree with themselves, the longer it takes them to give a final estimate. This suggests that two very different initial estimates cause a conflict that people tend to resolve by hedging against large errors (i.e., combining estimates). Second, there were no clear response time differences between the different strategies, except for when participants presumably started from scratch. The reason could be that only this strategy involves the retrieval of new knowledge, whereas all other strategies operate with the initial estimates as input.

The fact that we did not find differences in response time between the different strategies—with the exception of starting from scratch—raises the following possibility: The longer response times for final estimates that deviated more from an equal-

[6] Preliminary modeling using maximum likelihood estimation showed that including further fixed interaction effects (other than between logAD12 × Δ) or adding random effects for condition, logAD12, or Δ to participants, items, or both did not improve the model fit as indicated by the BIC. Removing the indicator variables for the dialectical conditions improved the model fit, but because those experimental manipulations were theoretically motivated, we still estimated the slopes for the respective indicator variables in the Bayesian regression model.

[7] We estimated three separate models because a single model trying to account for all interactions between the two indicator variables for the conditions and the three indicator variables for the revision strategies would have been overly complex.

weight strategy (see results for first regression model) may have been driven by the longer time it took participants when they started from scratch. Indeed, when we re-ran the first regression model and excluded the trials in which the final estimate was outside the range, a reliable effect of the deviation (i.e., Δ) was no longer found (0.08), HDI [−0.20, 0.32]; however, larger disagreements (i.e., logAD12) still predicted longer response times (0.13), HDI [0.08, 0.17].

General Discussion

The emerging field of research on the "crowd within" shows that people can, to some extent, simulate the "wisdom of crowds" within their own mind by averaging nonredundant estimates (e.g., Herzog & Hertwig, 2009; Vul & Pashler, 2008). Much less is known about whether people actually take advantage of the crowd within. Previous findings have indicated that people show consistency in how they deal with their conflicting estimates, with most people using the combining strategy (Müller-Trede, 2011), but important questions remained unanswered. For instance, will any repeated probing of the same issue trigger the use of the combining strategy? Or is combining more likely to happen when people explicitly take different perspectives on the same issue using a dialectical technique? Relatedly, is the use of the combining strategy rooted in people becoming cognizant of different views—or in the need to hedge one's risks?

This study offers answers to these questions: First, consistent with the diversity-inclusion hypothesis, people were more likely to combine their estimates when they actively contradicted themselves through dialectical bootstrapping (in this case, using the consider-the-opposite instruction as one possible dialectical technique) than when they simply estimated anew (reliability condition). Second, consistent with the hedging hypothesis, people were more likely to combine their estimates the more they disagreed with themselves; that is, people seem to hedge the risk of choosing the wrong estimate. Third, people who actively contradicted themselves were more inclined to combine (i.e., to place more equal weight on both opinions) than were people in the reliability condition—even with control for the magnitude of the disagreement. Thus, the support for the diversity-inclusion hypothesis is not confounded by the larger disagreements observed among people who applied dialectical bootstrapping.

Like Müller-Trede (2011), we found that participants who used a dialectical technique were consistent in how they arrived at their final estimate: Most of them tended to combine their estimates. Moreover, we found that the tendency to combine conflicting estimates was more pronounced both when participants actively contradicted themselves through dialectical bootstrapping (as opposed to merely making a new estimate; i.e., diversity-inclusion hypothesis) and as a function of the size of the disagreement within oneself (i.e., hedging hypothesis).

The finding that people show consistency in how they arrive at their final estimates (as both our Cronbach's α analyses and those by Müller-Trede, 2011, show) could have meant that people categorically combine whenever they are confronted with two conflicting self-generated estimates. Our findings rule out this interpretation, however. People are sensitive to how they generated their second estimate (diversity-inclusion hypothesis) and to how much they disagreed with themselves (hedging hypothesis) and thus do not approach every question with the same, preset strategy.

Using the Crowd Within and Dialectical Bootstrapping

To combine or not to combine—what should people do? It is a good strategy to reject combining in favor of choosing if (a) there are large differences in the accuracy of the estimates, (b) the errors between the estimates are similar, and (c) the probability of identifying the more accurate of the two estimates is high (Soll & Larrick, 2009). As our results show, however, these conditions barely apply to the crowd-within context: Errors of first and second estimates are similar in magnitude; errors are not redundant (especially with a dialectical technique); and people's ability to infer which estimate is more accurate is limited. Therefore, combining is likely to be the better strategy in the crowd-within context.

Of course, combining through dialectical bootstrapping will not always be superior to choosing. For example, first estimates of project completion times are notoriously optimistic and improve with debiasing (Buehler, Griffin, & Peetz, 2010). Therefore, choosing the second, and probably more realistic, estimate of the completion time will likely be superior to combining estimates. In many environments, however, it is difficult to figure out whether the conditions under which choosing pays are met. The three key variables—accuracy, error redundancy, and skill in picking the more accurate estimate—are typically unknown at the time of judgment and thus need to be gauged. Yet decision makers typically receive poor feedback from their environment on these parameters—especially about how correlated errors are (Larrick & Soll, 2006; Soll & Larrick, 2009). Consequently, the error-prone process of estimating the environmental parameters adds an additional layer of uncertainty, again favoring combining and dialectical bootstrapping over identifying the better estimate.

When evoking the crowd within, should a judge use a dialectical technique or simply ask herself again (as in the reliability condition)? We suggest that using a dialectical technique is a good default strategy for the following reason: Combining a first estimate with a dialectical estimate delivers either (a) superior averaging gains as compared with combining the first estimate with a second, nondialectical estimate or (b) at least the same averaging gains (Herzog & Hertwig, 2013; White & Antonakis, 2013). In either case, however, people using a dialectical technique appear to be more willing to combine their two estimates to produce a final estimate (as we showed in our study; i.e., diversity-inclusion hypothesis). Thus, people are more likely to realize potential averaging gains when using a dialectical technique—whether those gains are higher with a dialectical technique or the same.

Connections to Other Research

Advice taking. Research exploring how and when people use their crowd within has similarities with research on how people use others' advice (Bonaccio & Dalal, 2006). One robust finding in that literature is that revision strategies are egocentric (Bonaccio & Dalal, 2006; Soll & Larrick, 2009; Yaniv & Milyavsky, 2007). For example, when people are offered advice by a single advisor, they often completely ignore it and cling to their initial estimate (Soll & Larrick, 2009). When offered more than one piece of advice, they

incorporate advice similar to their own initial estimate but ignore inconsistent advice (Yaniv, 2004; Yaniv & Milyavsky, 2007).

Several explanations have been offered for the egocentric discounting of advice (Larrick, Mannes, & Soll, 2012). For example, it has been proposed that people hold the egocentric belief that they are more accurate than their advisors (Harvey & Harries, 2004), that they see their own estimates as more "objective" than those of their advisors (Liberman, Minson, Bryan, & Ross, 2012), or that they trust their own estimates more because they have introspective access to the reasoning that produced them (Yaniv & Kleinberger, 2000), but see Soll and Mannes (2011). Although those explanations differ greatly, most explain the discounting of advice in terms of differences that emerge because the advisor and the receiver are not the same person (e.g., information asymmetries or motivated cognition in social comparisons). Consequently, such explanations cannot be easily applied to the crowd-within context, which involves only one person and thus no information asymmetries or interpersonal processes. Some explanations are applicable to both domains, however. For example, incorrect intuitions about the benefits of averaging (Larrick & Soll, 2006; Soll, 1999) can be expected to hinder the use of combining in both advice taking and the crowd-within context. Furthermore, the hedging hypothesis we proposed should also be applicable to advice taking. Despite some theoretical crossover between the advice-taking and crowd-within literature, there is a need to develop explanations that are genuinely targeted at the crowd-within context, such as the diversity-inclusion hypothesis proposed and tested here.

Advice taking and confidence. People are more likely to use advice that is delivered with confidence, and receiving advice increases people's confidence in their revised estimates (see, e.g., the studies reviewed in Bonaccio & Dalal, 2006, pp. 132–133). It would be interesting to explore how people's confidence in the accuracy of their first and second estimates influences the way they take advantage of the crowd within.[8] Based on the hedging hypothesis, it might be predicted that people are more likely to combine the two estimates if the associated confidences are equal in size—that is, to hedge the risk of betting too much on one of the estimates. By contrast, when one estimate instills much more confidence than the other, this asymmetry may prompt people to choose rather than combine (see Hertwig, 2012; Koriat, 2012, for the related notion of maximum-confidence slating). Another interesting question is whether people have more confidence in their final estimate than in their first and second estimate, in the same way as they have higher confidence in their revised estimates after receiving advice.

Sensory integration of conflicting perceptual signals. The crowd within poses a challenge similar to that posed by conflicting perceptual signals.[9] People sometimes receive conflicting environmental information from different modalities (e.g., visual, auditory, and haptic) and need to integrate this information into a stable percept that can guide behavior (Berniker & Körding, 2011; Ernst & Bülthoff, 2004). Although some processes in perception follow the winner-takes-it-all principle (Sterzer, Kleinschmidt, & Rees, 2009), conflicting inputs from different modalities seem to be combined rather than selected between (Ernst & Bülthoff, 2004). From an optimality perspective, it follows that "[if] the goal is to come up with the most reliable (unbiased) estimate, then the variance of the final estimate should be reduced as much as possible" (Ernst & Bülthoff, 2004, p. 165). A perceiver can achieve minimal variance by averaging the conflicting perceptual signals and weighting them according to their reliability (i.e., accuracy). People's actual behavior in perceptual tasks is astonishingly close to such optimal behavior (Berniker & Körding, 2011; Landy, Banks, & Knill, 2011). However, optimal combination is not in fact necessary for good performance—in either cognitive or perceptual domains. Non-optimal combination of perceptual signals is quite accurate and superior to using just one perceptual modality (Oruç, Maloney, & Landy, 2003). Relatedly, the exact combination weights in advice taking are not crucial as long as they are not extreme (Soll & Larrick, 2009)—in the same way as the exact combination weights are not crucial when combining informational cues (Dawes, 1979).

Conclusions

A sole individual can boost his or her performance by simulating what would have appalled the French writer Le Bon (1895/1995): a crowd of diverse opinions within his or her own mind. Do people average across their diverse opinions and thus benefit from the crowd-within effect even when not instructed to do so? Our goal was to answer this question. When nudged with a dialectical bootstrapping technique (here, the consider-the-opposite instruction), people can create judgmental diversity (as shown previously) and, as we observed, they are indeed inclined to exploit this diversity by creating a mental "error portfolio" and combining their conflicting opinions.

[8] We thank the action editor and an anonymous reviewer for raising this point.

[9] We thank Ed Vul for pointing this out.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., . . . Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6,* 130–147. doi:10.1037/1076-898X.6.2.130

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic. doi:10.1007/978-0-306-47630-3_19

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes [Software]. Retrieved from http://CRAN.R-project.org/package=lme4

Berniker, M., & Körding, K. P. (2011). Bayesian approaches to sensory integration for motor control. *Wiley Interdisciplinary Reviews: Cognitive Science, 2,* 419–428. doi:10.1002/wcs.125

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101,* 127–151. doi:10.1016/j.obhdp.2006.07.001

Buehler, R., Griffin, D., & Peetz, J. (2010). The planning fallacy: Cognitive, motivational, and social origins. In M. P. Zanna & Olsen (Eds.), *Advances in experimental social psychology* (Vol. 43, pp. 1–62). San Diego, CA: Academic Press. doi:10.1016/S0065-2601(10)43001-4

Carlyle, T. (Ed.). (1855). *Oliver Cromwell's letters and speeches*. New York, NY: Harper.

Central Intelligence Agency. (2008). *The world factbook*. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. doi:10.1037/0003-066X.34.7.571

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi:10.1177/1745691611406920

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8,* 162–169. doi:10.1016/j.tics.2004.02.002

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*. Retrieved from http://www.jstatsoft.org/v33/i02/paper

Harvey, N., & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting, 20,* 391–409. doi:10.1016/j.ijforecast.2003.09.012

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science, 336,* 303–304. doi:10.1126/science.1221403

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20,* 231–237. doi:10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2013). The crowd-within and the benefits of dialectical bootstrapping: A reply to White and Antonakis (2013). *Psychological Science, 24,* 117–119. doi:10.1177/0956797612457399

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: Selecting among forecasts and their combinations. *International Journal of Forecasting, 21,* 15–24. doi:10.1016/j.ijforecast.2004.05.002

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1068–1074. doi:10.1037/a0019694

Koriat, A. (2012). When are two heads better than one and why? *Science, 336,* 360–362. doi:10.1126/science.1216549

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300. doi:10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. doi:10.1177/1745691611406925

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142,* 573–603. doi:10.1037/a0029146

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-observer models of cue integration. In J. Trommershäuser, K. P. Körding, & M. S. Landy (Eds.), *Sensory cue integration* (pp. 5–29). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195387247.003.0001

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Oxford, England: Blackwell. doi:10.1002/9780470752937.ch16

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (pp. 227–242). New York, NY: Psychology Press.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52,* 111–127. doi:10.1287/mnsc.1050.0459

Le Bon, G. (1995). *The crowd.* New Brunswick, NJ: Transaction. (Original work published 1895)

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads". *Journal of Experimental Social Psychology, 48,* 507–512. doi:10.1016/j.jesp.2011.10.016

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3,* 112–115. doi:10.1111/j.2041-210X.2011.00131.x

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231–1243. doi:10.1037/0022-3514.47.6.1231

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software, 42*. Retrieved from http://www.jstatsoft.org/v42/i09/paper

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making, 6,* 283–294. Retrieved from http://journal.sjdm.org/11/101217a/jdm101217a.pdf

Oruç, İ., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research, 43,* 2451–2468. doi:10.1016/S0042-6989(03)00435-8

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Plummer, M. (2012). *JAGS version 3.3.0 user manual*. Retrieved from http://mcmc-jags.sourceforge.net/

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology, 55,* 191–197. doi:10.1016/j.jmp.2010.10.002

Santos, A. (2009). *How many licks? Or, how to estimate damn near anything*. Philadelphia, PA: Running Press.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology, 38,* 317–346. doi:10.1006/cogp.1998.0699

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 780–805. doi:10.1037/a0015145

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting, 27,* 81–102. doi:10.1016/j.ijforecast.2010.05.003

Sterzer, P., Kleinschmidt, A., & Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Sciences, 13,* 310–318. doi:10.1016/j.tics.2009.04.006

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

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*. Garden City, NY: Doubleday.

Vul, E. (2008). *Crowd within*. Retrieved from http://web.archive.org/web/20080828123805/http://www.edvul.com/crowdwithin.php

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

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

White, C. M., & Antonakis, J. (2013). Quantifying accuracy improvement in sets of pooled judgments: Does dialectical bootstrapping work? *Psychological Science, 24,* 115–116. doi:10.1177/0956797612449174

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

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93,* 1–13. doi:10.1016/j.obhdp.2003.08.002

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83,* 260–281. doi:10.1006/obhd.2000.2909

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes, 103,* 104–120. doi:10.1016/j.obhdp.2006.05.006

Appendix A

Questions Used in the Study and Correct Answers (in %)

Question	Answer
The area of the USA is what percent of the area of the Pacific Ocean?	6
What percent of the world's population lives in either China, India, or the European Union?	44
What percent of the world's airports are in the United States?	30
What percent of the world's roads are in India?	11
What percent of the world's countries have a higher fertility rate than the United States?	58
What percent of the world's telephone lines are in China, USA, or the European Union?	72
Saudi Arabia consumes what percentage of the oil it produces?	19
What percentage of the world's countries have a higher life expectancy than the United States?	20
What percent of the earth's surface is covered by water?	71
What percent of the worldwide land mass is not used for agriculture?	82
What percent of the world's population is between 15 and 64 years old?	65
What percent of the world's population is Christian?	33
What percent of the world's population speaks Mandarin Chinese as a first language?	13
What percent of the world's population aged 15 years or older can read and write?	82
What percent of the worldwide gross domestic product (GDP) comes from the service sector?	64
What percent of the worldwide labor force works in the agricultural sector?	40
What percent of the worldwide income does the richest 10% of households earn?	30
What percent of the worldwide gross domestic product (GDP) is re-invested ("gross fixed investment")?	23
What percent of the goods exported worldwide are mineral fuels (including oil, coal, gas, and refined products)?	14
What percent of the worldwide gross domestic product (GDP) is used for the military (military expenditure)?	2

Note. The first eight items (as used in Vul & Pashler, 2008) were taken from "Crowd Within," by E. Vul, 2008 (http://web.archive.org/web/20080828123805/http:/www.edvul.com/crowdwithin.php). Copyright 2008 by E. Vul. The remaining 12 items were created based on facts reported in *The World Factbook* (Central Intelligence Agency, 2008). All questions were translated into German for the studies.

(*Appendices continue*)

Appendix B

Description of Bayesian Statistics

Below we detail the statistical models. Except where noted, all models were implemented as Markov Chain Monte Carlo (MCMC) models in *Just Another Gibbs Sampler* (JAGS; Plummer, 2012) with three chains (using different starting values) with 50,000 samples each (using another 50,000 samples for burn-in). Because thinning lowers the precision of the posterior estimates (Link & Eaton, 2012), no thinning was used.

Note that because we used MCMC, the mode of a posterior distribution of a linear contrast (e.g., the two dialectical conditions vs. the reliability condition, as reported in the main text) need not coincide with a linear contrast value that is calculated using the modes of the posterior distributions of the individual conditions (as presented in Table 1).

Categorical Data

For binary categorical data (e.g., whether participants scored a "success" or not), we modeled the success probability θ based on the number of k observed successes out of n participants using the binomial likelihood function—assuming a uniform prior on θ (Kruschke, 2011b). For categorical data with more than two categories (e.g., classifying participants to strategies A, B, and C), we modeled the probabilities θ_i of belonging to the ith category based on the observed category counts using the multinomial function—assuming a uniform Dirichlet prior on the θ_is; those posteriors were directly approximated with the R function "MCmultinomdirichlet" from the R package "MCMCpack" (Martin, Quinn, & Park, 2011) using 1,000,000 samples.

Continuous Distributions

Because of fat tails and occasional outliers, we used t distributions to model continuous data (extending the approach in Kruschke, 2013, to three groups). Like a normal distribution, a t distribution has a mean and a standard deviation; it additionally has a third parameter that indicates the amount of kurtosis in the data (normality parameter ν). Lower values of ν indicate fatter tails; as ν approaches positive infinity, the t distribution becomes the normal distribution. Unlike a normal distribution, a t distribution can accommodate fat tails and outliers and thus leads to more robust inferences; furthermore, if the data happen to be normal, the t distribution will "mimic" this. We adopted Kruschke's (2013) vague priors for the means, the standard deviations, and ν. We estimated separate means and standard deviations for each experimental group, but only one ν across all conditions. Note that the t distribution is used as a model for the data (and not as a sampling distribution from which p values are derived in a frequentist approach).

Proportions

We used a hierarchical model to estimate group-level proportions (i.e., the proportion of times a "typical" participant scores a "success" in n trials), separately for the experimental conditions (see Kruschke, 2011b, pp. 219–224). Each participant was assumed to have an individual probability θ_i of scoring a success (e.g., bracketing); each θ_i was estimated based on the number of k observed successes out of n trials using the binomial distribution. The θ_is were estimated hierarchically by assuming that each θ_i comes from a group-level distribution of θs following a beta distribution (which can range from 0 to 1). We parameterized the beta distributions in terms of a mean and a sample size (instead of the two shape parameters α and β). The prior for the mean followed a uniform distribution between 0 and 1; the prior for the sample size followed a gamma distribution (with the parameters for shape and rate set to 1 and 0.1, respectively, which implies a vague prior with both a mean and a standard deviation of 10; see Kruschke, 2011b, p. 211). For this model, 950,000 burn-in samples per chain were used.

Mixed-Effects Linear Models

We fitted the mixed-effects linear regression models (Baayen et al., 2008) using the R package "MCMCglmm" (Hadfield, 2010) and used the software default vague (improper) priors. We first explored reasonable model specifications with maximum likelihood estimation using the function "lmer" from the R package "lme4" (Bates, Mächler, & Bolker, 2011) to identify the model to be estimated with MCMCglmm.

Received September 7, 2012
Revision received June 14, 2013
Accepted June 17, 2013 ■