Judgmental aggregation strategies depend on whether the self is involved

Jack B. Soll\textsuperscript{a,}\textsuperscript{*}, Albert E. Mannes\textsuperscript{b}

\textsuperscript{a} Fuqua School of Business, Duke University, United States
\textsuperscript{b} Tepper School of Business, Carnegie Mellon University, United States

Abstract

We report the results of a novel experiment that addresses two unresolved questions in the judgmental forecasting literature. First, how does combining the estimates of others differ from revising one’s own estimate based on the judgment of another? The experiment found that participants often ignored advice when revising an estimate but averaged estimates when combining. This was true despite receiving identical feedback about the accuracy of past judgments. Second, why do people consistently tend to overweight their own opinions at the expense of profitable advice? We compared two prominent explanations for this, differential access to reasons and egocentric beliefs, and found that neither adequately accounts for the overweighting of the self. Finally, echoing past research, we find that averaging opinions is often advantageous, but that choosing a single judge can perform well in certain predictable situations.

Keywords: Advice taking; Averaging judgments; Combining opinions; Information integration; Judgmental forecasting

1. Introduction

A robust phenomenon in the study of human judgment is that people tend to overweight their existing beliefs in the face of new information. This basic finding has been demonstrated across a range of different types of judgments. People are conservative in updating judgments of probability (Edwards, 1968), are unduly influenced by first impressions in their judgments of others (Asch, 1946; Kelley, 1950), and persevere in their beliefs in the face of contradictory evidence (Nisbett & Ross, 1980). In the context of judgmental forecasting, people change their minds more slowly than they should, as evidenced by the accuracy of their final judgments. For example, Lim and O’Connor (1995) examined the way in which people use model-based forecasts to supplement their own. They found that people tenaciously favor their own judgment over valid models to the detriment of accuracy, a tendency that worsened over time.

A recent area of interest has been the way in which people revise opinions after learning the opinion of
an advisor (e.g., Bonaccio & Dalal, 2006; Sniezek & Buckley, 1995). A common finding for quantitative estimates is that, on average, people tend to adjust about 20%–30% of the way toward advice (Harvey & Fischer, 1997; Soll & Larrick, 2009; Yaniv, 2004b), a phenomenon that Yaniv and Kleinberger (2000) labeled “egocentric discounting”. This corresponds to an overestimation of the self, because a simple average of opinions often leads to greater accuracy (Mannes, 2009; Soll & Larrick, 2009; Yaniv, 2004a). Averaging works because estimates often bracket (i.e., fall on opposite sides of) the truth, causing positive and negative errors to cancel out (Larrick & Soll, 2006). Although the finding of 20%–30% weight on advice is a common one, it is important to point out that this is an aggregate result that may not accurately reflect item-level behavior. Soll and Larrick (2009) showed that people often either choose one of the two answers (often their own) or use an equal-weighted average. In their studies, the 30% mean adjustment reflected a pattern of frequently ignoring advice, sometimes averaging, and occasionally fully accepting advice. Participants paid a price for this, as they would have been more accurate had they consistently averaged.

A related stream of research has focused on how people combine the opinions of others (Birnbaum & Stegner, 1979; Budescu, 2006; Harvey, Harries, & Fischer, 2000; Yaniv, 1997). David Budescu and his colleagues (Budescu, 2006; Budescu & Rantilla, 2000; Budescu, Rantilla, Yu, & Karelitz, 2003; Budescu & Yu, 2007) have tested the fit of several models of judgmental aggregation for combining the forecasts of expert advisors, including choosing the best advisor, simple averaging, and weighted averaging. The judgments of most participants in their studies were best fit by a weighted averaging model, where the weights depended on the advisors’ forecasting skills and access to information (Budescu, 2006). This differs from Soll and Larrick’s (2009) conclusion that people often choose (i.e., use weights of 0 and 1) when they revise their own opinion. Formally, the two tasks are identical, in the sense that if beliefs about statistical relationships such as relative accuracy and intercorrelation among judges are the same, then a formal model would prescribe the same aggregation rule. A puzzle, therefore, is to explain the differences in weighting for the two tasks.

One potential explanation is that beliefs in the two tasks are not the same. For instance, the high rate of ignoring advice in opinion revision might reflect a belief that the advisor is substantially less accurate than oneself (Harvey & Harries, 2004). Alternatively, Yaniv and Kleinberger (2000) proposed that egocentric discounting of advice arises because people have access to the reasons for their own judgment but not to the reasons of others. They built on support theory (Tversky & Koehler, 1994), suggesting that estimation is a process of weighting possible answers based on the supporting evidence. Because one’s own reasons are often richer and more salient than those of others, others’ opinions attract less weight in the revision process (Yaniv, 2004b; Yaniv & Kleinberger, 2000).

In addition to these two explanations, studies of opinion revision and combining others’ estimates have differed procedurally, which may also account for the differences in results. For instance, many studies of opinion revision have involved updating one’s own opinion based on the opinion of a single advisor (for exceptions see Mannes, 2009; Yaniv & Milyavsky, 2007), whereas studies of combining others’ opinions have typically involved several opinions being combined. The two paradigms have also used different methods to reach their conclusions. Whereas most studies of opinion revision have analyzed means or distributions of weights, studies of combining others’ opinions have used model fitting techniques to compare the fits of different strategies. These approaches are all valid and informative, but differences in procedures and analyses make it difficult to draw firm conclusions about the sources of any differences in results.

A chief goal of this paper is to compare the tasks of revising and combining others’ opinions in a single experiment. This will allow us to determine if and how judgmental aggregation strategies differ depending on whether the self is involved. To foreshadow one result, we show that people weight a judge’s opinion more when they themselves are that judge, holding constant the available information about the accuracy of the judges. Given this difference, we then consider potential explanations. To what extent can this difference be explained by biased beliefs about relative accuracy, or by differential access to reasons? Finally, we examine accuracy in both tasks relative to benchmark strategies such as consistently averaging opinions. Although past research has generally been supportive of averag-
ing, its success depends on the nature of the environment (Soll & Larrick, 2009). We therefore manipulate key features of the environment in order to identify the conditions under which people perform well or poorly on the two tasks.

The paper is organized as follows. We first provide additional background on potential explanations for the egocentric discounting of advice. Next, given that participants could often have improved had they averaged consistently, we review the logic for why averaging is an effective strategy, and discuss the boundary conditions of averaging. We then describe an experiment that compares the tasks of opinion revision and combining others’ opinions. The results show the ways in which weighting strategies in the two tasks differ, and also provide insight into the viability of competing explanations for these differences. We conclude with a discussion of the implications for both psychology and practice.

2. Egocentric discounting of advice

The most prominent explanations for egocentric advice discounting are differential access to reasons and biased beliefs about ability. The first explanation asserts that people weight opinions in relation to the strength of the supporting evidence. When people have access to their own reasons but not the reasons of others, the balance of evidence will tend to favor the initial answer, leading to the discounting of others’ opinions (Yaniv, 2004b; Yaniv & Kleinberger, 2000). More generally, the mechanism of accessibility has received substantial empirical support as an explanation for judgmental biases, such as the anchoring effect (Chapman & Johnson, 2002; Mussweiler & Strack, 1999). The starting premise is that judgment is based on the sample of evidence that comes to mind. When an anchor (which can even be a randomly generated number) is presented before a person independently considers a quantitative problem, processes of associative memory cause anchor-consistent evidence to be disproportionately sampled and influential on judgment. In advice taking, an accessibility model would predict that an individual’s answers would be closer to the advice when it comes before the judge has formed an opinion as opposed to afterwards. This is precisely what happens (Koehler & Beauregard, 2006; Sniezek & Buckley, 1995). In most studies of advice taking, however, the judge first forms an initial opinion independently, then receives advice. An accessibility explanation might state that the judge retrieves a coherent and consistent set of evidence favoring the initial answer, and perhaps even continues to build on this after the answer has been reported (Mussweiler & Strack, 1999). According to this account, advice would face two disadvantages: the recruitment of support for an advisor’s answer may be inhibited by processes of memory, motivation, or biased interpretation (Klayman, 1995), and the judge generally lacks access to the advisor’s thoughts (Yaniv & Kleinberger, 2000).

An alternative explanation is that people weight judges as a function of perceived expertise. Others’ advice will be egocentrically discounted if beliefs about relative accuracy are biased in favor of the self (Harvey & Harries, 2004). According to this view, people attend to cues to expertise such as experience, past performance, and reputation. One source of bias is that people may be motivated to interpret ambiguous evidence for their skills in a self-serving manner (Dunning, Meyerowitz, & Holzberg, 1989). Bias may also arise from an asymmetry in information about the self and others (Burson, Larrick, & Klayman, 2006; Moore & Small, 2007). For example, a person may be impressed by the cogency of his or her own thoughts, but may have substantially less information about whether others’ thoughts are cogent (Mojzisch, Grouneva, & Schulz-Hardt, in press; Van Swol, Savadari, & Sniezek, 2003). If most people have cogent thoughts, then most people would also infer that they are more accurate than average, and a systematic overweighting of the self would follow as a consequence. In this way, differential access to reasons can have an indirect impact on weighting, by feeding into judgments of relative expertise.

Clearly, the two proposed explanations need not be mutually exclusive. In fact, Yaniv and Kleinberger (2000) suggested that people are sensitive not only to the balance of available reasons, but also to the perceived quality of the judges, based on past performance. To accommodate this, the reason-based model could be construed such that support for a given answer derives from both factual knowledge and the perceived expertise of judges who provided that answer. Given that introspection about reasons can influence beliefs about expertise, and also that beliefs about expertise can provide support in a weighting
process, the two explanations may not be as distinct as they appear on the surface.

In addition to these explanations, it has also been suggested that people anchor on their own answer and then adjust toward advice as a function of perceived expertise or credibility (Lim & O’Connor, 1995). Since adjustment from an internally generated anchor is typically insufficient (Epley & Gilovich, 2004), this would lead to an over weighting of the self. This sort of adjustment process seems unlikely in this context, given that people rarely adjust 30% of the way toward advice. In fact, they often do not adjust at all, occasionally average, and sometimes adjust all the way, leading to a tri-modal distribution of adjustments (Soll & Larrick, 2009). If advice taking were a process of adjusting from an anchor, albeit insufficiently, then the distribution of weights should be less lumpy than what is observed, with far fewer cases of not adjusting at all (see Harvey & Harries, 2004, for an alternative approach to ruling out insufficient adjustment as a mechanism).

3. The benefits of averaging

One of the most effective ways of combining quantitative opinions is simply to average them. Averaging opinions is remarkably robust, and compares favorably to more complex methods of aggregation (Armstrong, 2001; Clemen, 1989; Goodwin & Wright, 1998). Yet, the often stellar performance of averaging continues to surprise people. In fact, early scientists did not trust averaging—it was not until the late nineteenth century that scientists routinely averaged multiple measurements of the same object. Prior to that, they would often choose the single measurement that they believed to be the most accurate (Stigler, 1999). Many laypeople have the same intuition as the early scientists, believing that the average opinion of a group will be about as accurate as that of its average member (Larrick & Soll, 2006). The substantial improvement that can be achieved by combining opinions is so surprising that two popular books have been written on the subject (Page, 2007; Surowiecki, 2004).

To see why averaging works, consider a simple example of two economists, an optimist and a pessimist, estimating GNP growth over the next year. Suppose that the optimist guesses that the economy will grow by 10%, whereas the pessimist says −2%. Suppose that the realized value turns out to be 6%, which falls strictly between the two estimates. When the estimates bracket the truth in this fashion, the absolute deviation (AD) of the average is always less than the average AD of the individual estimates. In this example, \( AD_{\text{averaging}} = 2 \) and average \( AD = 6 \). Next, consider what happens when the estimates do not bracket the truth. If GNP growth turns out to be 12% instead, then \( AD_{\text{averaging}} = \) average \( AD = 8 \). Putting the two results together, over a series of forecasts averaging will perform at least as well as the average judge, and will outperform the average judge as long as there is at least one instance of bracketing. Empirically, bracketing rates for two judges are typically in the range of 30%–50%, which allows for substantial benefits from averaging (Soll & Larrick, 2009). The chance of bracketing also increases with the number of judges. In general, the accuracy of a composite increases at a diminishing rate; typically, the full benefit of averaging is realized with 8–20 opinions (Hogarth, 1978).

Of course, averaging may not always be effective, such as when there are large differences in forecasting skill. In general, the effectiveness of a judgmental strategy is contingent on features of the environment (Payne, Bettman, & Johnson, 1993). In the case of quantity estimates, the PAR (probability, accuracy, redundancy) model has been developed to compare the accuracy of equal-weight averaging and choosing a single judge (Soll & Larrick, 2009). There are several reasons for focusing on these two strategies. First, these are the strategies that people most frequently employ in revising their own judgments (Soll & Larrick, 2009). Second, the strategies map onto two opposing views in the judgment and decision making literature on how to best use multiple pieces of information—equally weight the relevant cues or attributes (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975), or use the “take-the-best” heuristic and rely on the single most valid cue (Gigerenzer & Goldstein, 1996). Third, in many judgmental situations there is insufficient data to reliably derive optimal weights, in which case a sound strategy is to identify the most valid pieces of information and weight them equally. In many realistic situations, these unit weights can perform as well as or better than attempts to estimate the weights with formal methods such as regression (Dawes, 1979; Einhorn & Hogarth, 1975).
The PAR model identifies the better strategy based on three parameters: the probability of identifying the better judge, the relative expertise of the two judges, and the frequency with which the judges’ answers bracket the truth. The relationships among the parameters are shown in Fig. 1. The accuracy ratio $A$, plotted on the x-axis, is defined as the ratio of the MADs (mean absolute deviations) of the two judges, higher over lower. Higher levels of $A$ correspond to larger differences in the expertise of the two judges. The bracketing rate is plotted on the y-axis, and reflects the extent to which judges share a bias or have correlated errors. The “iso-accuracy” curve, so named because the expected performance of choosing and averaging is identical along the curve, divides the environmental space into regions that favor averaging and choosing. There are several iso-accuracy curves plotted in the figure, reflecting different probabilities of identifying the better judge. The shaded region above the thick curve shows environments in which averaging is the superior strategy when the more accurate judge can be identified with certainty (i.e., $p = 1$). As the probability declines, the iso-accuracy curve rotates downward, enlarging the space that favors averaging.

In general, averaging performs better than choosing when the probability of detecting the expert is low, differences in expertise are low, and bracketing is high (i.e., the redundancy in errors is low). As shown in the figure, averaging can still be effective with a high accuracy ratio if bracketing is also high. For example, at bracketing rates above 60%, averaging is the superior strategy even when the inferior judge is twice as inaccurate as the better judge and the ranking of the judges is known with certainty. Of course, in many situations the better judge cannot easily be identified (e.g. Baumann & Bonner, 2004; Littlepage, Robison, & Reddington, 1997), which increases the relative advantage of averaging over choosing a single judge. This example illustrates why empirical demonstrations of averaging have been so impressive: the strategy is robust to fairly large differences in expertise.

4. Experiment

We designed an experiment to examine the differences between revising one’s own opinion and combining the opinions of others. One of the key features of the design is that participants who combined the opinions of others were yoked to participants who revised, such that each yoked pair received the same feedback on historical performance and worked with the same set of estimates. If a given judge is weighted more heavily when the participant is that judge, this would be evidence of an egocentric bias in discounting advice. We also elicited judgments of relative expertise and confidence, in order to examine the extent to which biased perceptions of the self might account for advice discounting.

The task involved predicting the average points per game (APPG) of 30 men’s college basketball teams during the 2008–2009 season based on four cues (rebounds per game, steals per game, field-goal percentage, and 3-point percentage). The identities of the teams were kept anonymous, so that participants had to rely on the available cues in their predictions. There were two conditions in which participants could revise their estimates based on the answers of an advisor. Both revision groups provided independent opinions on the 30 games, saw feedback on a sample of games (including both judges’ estimates and the truth), and then revised their answers for 20 of the games based on their own answer and the advice. In the revise-with-cues condition, participants continued to see the cues while they revised, whereas in the revise-without-cues condition the answers were presented without the cues. We asked a third set of participants to predict the APPG based solely on the original estimates of two other students. In fact, these participants were yoked to those who revised without
the cues, such that the initial judgments of the self and advisor became, for this group, the opinions of Judge 1 and Judge 2. They also saw the same feedback as those in the revision condition, but did not see the cues during the combination phase of the study.

Since participants in the revise-without-cues and combine-others tasks had the same feedback information and combined the same estimates without the benefit of reasons, any differences between these two conditions would reflect either differences in perceptions of relative ability or differences in aggregation strategies. Obviously, differences between these two conditions cannot be due to reasons, because reasons were not available at the time of revision. We tested the reason-based explanation by including the revise-with-cues condition. When cues are present, the judge can assess the implications of those facts and develop a rule for combining them into a judgment. If people weight their own opinions more when they view the cues in the revision stage, this would be evidence that they weight answers according to their subjective impressions of the supporting evidence.

We used three environments in this experiment, denoted by Equal Accuracy (EA in Fig. 1), Self Better (SB), and Advisor Better (AB). These environments are crossed with the aforementioned tasks: opinion revision (with or without cues) and combining others’ opinions. In the Equal Accuracy environment, the advisor is simulated to be as accurate as the participant (over a set of 30 questions, as described below) and the bracketing rate is 50%. In this environment, averaging is expected to improve judgmental accuracy by about 30% compared to choosing one of the two judges. In both the Self Better and Advisor Better environments the bracketing rate is 30% and one judge is 70% less accurate than the other ($A = 1.7$), so they are both at the same point in Fig. 1. For both of these environments, averaging is approximately 30% less accurate than choosing the better judge with certainty. By creating three distinct environments, we can investigate how well people match their aggregation strategies to the environment for both tasks. Soll and Larrick (2009) showed that people do not adapt sufficiently when revising—they tend to choose themselves across environments, which leads them to underperform averaging in environments that favor either averaging or choosing the advisor. The current design allows us to extend this analysis to the task of combining others’ judgments.

In addition to examining the environment and the task, we also manipulated the evidence about judge expertise by varying the amount of feedback on each judge’s past performance. Some participants saw estimates and truth for 3 past trials, whereas others saw results for 10 trials. Sample performance can be thought of as a measure of true ability plus error. Since small samples have more error, they will also tend to produce greater absolute differences between the sample performances of the two judges. If people are appropriately sensitive to sample size, then small samples should lead to more regressive beliefs (toward equal accuracy) and more frequent averaging. On the other hand, if people incorrectly treat the sample as representative of true differences in predictive ability, then small samples should lead to perceptions of greater differences between the judges, and therefore more choosing.

To summarize, we have four objectives in this experiment. The first is to compare the aggregation strategies that people use when they revise versus when they combine the opinions of others. Our review of the extant literature suggests that people often choose when they revise but average when they combine others’ opinions. However, the two tasks have not previously been compared in the same experiment, so many factors could account for this apparent difference. By fixing the number of judges at two and controlling for feedback and beliefs, our design allows us to determine whether people really do approach these two tasks differently. Second, we seek to find evidence for or against competing explanations for the egocentric discounting of advice. We do this by comparing three task conditions: revise-with-cues, revise-without-cues, and combine-others. Third, by varying the amount of feedback, we examine whether people are appropriately sensitive to sample size when forming beliefs and aggregating opinions. If people treat sample performance as perfectly representative of true skill, then they are likely to choose more and average less with small samples of feedback, which is the opposite of what they should do to maximize accuracy. Fourth, in manipulating both the task and the environment, our design allows us to examine the conditions under which people perform well or poorly relative to benchmark strategies such as consistent averaging.
4.1. Method

4.1.1. Stimuli

Information on the criterion APPG and the four predictive cues was collected for men’s college basketball teams from seven large conferences, comprising 82 teams in the initial sample. The data were collected from a public website, www.cbssports.com, for the 2008–2009 season. Ten of the teams were randomly chosen for the training trials and 30 of the remainder were chosen for the test trials. All participants judged the same teams in the same order. Descriptive statistics for these variables are presented in Table 1.

4.1.2. Participants

One hundred and ninety-eight students (104 female, mean age of 21.1) from a private US university participated in the experiment for a guaranteed payment of $6 and a bonus based on performance ($M = 4.10, SD = 0.74).

4.1.3. Procedure

Participants were seated at private computer terminals in the university’s behavioral lab. Each terminal ran the three task conditions in the following sequence: revise-with-cues, revise-without-cues, and combine-others. The ordering of the latter two conditions was critical, because those in the combine-others condition were presented with the judgments of those in the revise-without-cues condition. In general, there were four phases of the procedure: training, initial judgments, environmental assessment, and evaluation. The instructions and procedure for each phase varied by task (see Table 2).

In the revise-with-cues condition, participants were told that they were to predict the APPG for various men’s college basketball teams over the 2008–2009 season based on four pieces of information (cues). On each trial, the same four cues were provided for the team, which was anonymous (e.g., Team 1, Team 2, etc.). For example, the first question informed participants that Team 1 had 38.4 rebounds per game, 6.8 steals per game, a 40.6% field-goal percentage, and a 29.6% 3-point percentage. Below this information, participants were asked to “Please enter your best estimate of this team’s average points per game (APPG): ______.”

Participants made their estimates and were provided with immediate feedback—namely, the team’s actual APPG and their absolute deviation—after which they proceeded to the next trial. Participants were given 10 training trials in order to familiarize them with the task. The purpose of the training trials was to calibrate participants on the scale of the criterion and to give them a sense (not mastery) of the cue-criterion correlations. Following their training, participants repeated the prediction task for 30 new teams, with two differences. First, they did not receive feedback on each trial; they learned their performance only at the end of the experiment. Second, they were told that they would be rewarded for the accuracy of their predictions for 20 of the 30 teams chosen at random. The participants were given instructions and tested on the calculation of MAD, the standard used for the evaluation of their performance (they earned a $1 bonus for achieving a MAD below 12, and an additional $1 for each 2-unit reduction in MAD, up to $5 maximum). The 30 teams were presented sequentially and anonymously (e.g., Team 1), together with their four statistics. Participants entered and confirmed their predictions on the computer, then moved on to the next team. They could not return to prior estimates and received no feedback on their performance while making their predictions.

After making their initial predictions, the participants were informed that they had the opportunity to revise the 20 estimates on which they were to be evaluated, based on the advice of another student who had completed the task previously. Before doing so, they were to review the accuracy of their initial predictions and those of their advisor. At this point they were shown a table with either 3 (small sample size) or 10 (large sample size) trials. For each trial, the team was listed (e.g., Team A) along with the team’s actual APPG, the participant’s initial prediction, and the advisor’s prediction. The advisors’ predictions were simulated based on the participants’ initial predictions and their environment (see Appendix). The table remained visible while the participant assessed the environment by answering five questions: (1) Who was more accurate (yourself or the other student)? (2) How confident are you in this choice (50%–100%)? (3) What is your MAD? (4) What is the MAD of the other student? (5) How often do you expect that you and the other student will make estimates that fall on opposite sides of the truth (0–20)? After answering these questions, the participants were reminded of the bonus structure.
Table 1
Descriptive statistics for the basketball prediction task.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Average points per game</td>
<td>71.7</td>
<td>6.3</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(criterion)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Rebounds per game</td>
<td>38.7</td>
<td>3.6</td>
<td>0.67**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Steals per game</td>
<td>6.8</td>
<td>1.3</td>
<td>0.48**</td>
<td>0.23*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Field-goal percentage</td>
<td>45.0</td>
<td>2.2</td>
<td>0.55**</td>
<td>0.09</td>
<td>0.15</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>5. 3-point percentage</td>
<td>35.2</td>
<td>2.5</td>
<td>0.27*</td>
<td>−0.21</td>
<td>−0.09</td>
<td>0.51**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

n = 82.
** p < 0.01.
* p < 0.05.

Table 2
Experimental tasks and procedures.

<table>
<thead>
<tr>
<th>Task</th>
<th>Training</th>
<th>Initial estimates</th>
<th>Environmental assessment</th>
<th>Final estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revise with cues</td>
<td>10 trials predicting APPG with authentic cues</td>
<td>30 trials predicting APPG with authentic cues</td>
<td>Indicate beliefs about self and advisor based on 3- or 10-trial feedback</td>
<td>20 trials revising initial APPG estimates with cues and advice</td>
</tr>
<tr>
<td>Revise without cues</td>
<td>10 trials predicting APPG with authentic cues</td>
<td>30 trials predicting APPG with authentic cues</td>
<td>Indicate beliefs about self and advisor based on 3- or 10-trial feedback^a</td>
<td>20 trials revising initial APPG estimates with advice but without cues^a</td>
</tr>
<tr>
<td>Combine others</td>
<td>10 trials predicting an unrelated criterion with generic cues</td>
<td>30 trials predicting an unrelated criterion with generic cues</td>
<td>Indicate beliefs about two judges based on 3- or 10-trial feedback^a</td>
<td>20 trials combining two estimates of APPG without cues^a</td>
</tr>
</tbody>
</table>

n = 66 per task.
^a Yoked stimuli.

and the task. The 20 teams (which did not overlap with the teams appearing in the feedback trials) were then presented sequentially, together with their four cues, the participant’s initial prediction, and his or her advisor’s prediction. The participants entered and confirmed their final prediction of the APPG for each team and then moved to the next team. They could not return to prior estimates and received no feedback on their performance.

The revise-without-cues condition was identical to revising with cues in all respects but one: the 20 teams in the revision phase were still presented with the participant’s initial prediction, and that of the advisor, but without the four cues. The participants were aware that the cues would not be available before they started this phase of the experiment.

The participants in the combine-others condition had the task of combining the opinions of two other students, rather than of revising their own. Thus, they did not need training on the four cues, nor did they need to make 30 predictions of their own. Nevertheless, to keep the experimental conditions as similar as possible (in terms of time, fatigue, etc.), these participants were told that there were two unrelated parts of the experiment: a prediction task and a task assessing how they use information provided by others. For the prediction task, the participants were to produce estimates of a criterion based on four pieces of information. They would have 10 practice trials with feedback and 30 evaluation trials without feedback. In reality, the stimuli and procedure were identical to those in the two revision conditions, except that no context was provided (namely, about basketball), the cues and criterion were disguised (e.g., cue 1), and the criterion values (the actual APPG) were halved, to change the scaling between this and their second task. They were evaluated based on their MAD scores at the end of the 30 trials, but only qualitatively (e.g., exceptional, very good, fair, etc.), not financially.

After this unrelated prediction task, the participants in the combine-others condition were introduced to the task on information use. They were informed
they were to predict the APPG for 20 men's college basketball teams based only on advice provided by two students on a prior occasion. Before doing so, they were to review the accuracy of these two students. At this point they saw a table with either 3 (small sample size) or 10 (large sample size) trials. For each trial, the team was listed (e.g., Team A), together with the team’s actual APPG and the predictions of the two students. In reality, the predictions of Judge 1 were the initial predictions of the participant to whom they were yoked from the prior condition (revise-without-cues), and the predictions of Judge 2 were those of the prior participant’s advisor. The table remained visible while the participant answered five questions: (1) Who was more accurate (Judge 1 or Judge 2)? (2) How confident are you in this choice (50%–100%)? (3) What is the MAD of Judge 1? (4) What is the MAD of Judge 2? (5) How often do you expect the two students to make estimates that fall on opposite sides of the truth (0–20)? After answering these questions, the participants were introduced to the financial bonus structure and task. The 20 teams were presented sequentially, together with the predictions of the two advisors. No other information was available about the teams. The participants entered and confirmed their prediction of the APPG and then moved on to the next team. They could not return to prior estimates and received no feedback on their performance. At the conclusion of the experiment, the participants learned their MADs on the 20 evaluation trials and their bonuses, after which they were debriefed and dismissed.

4.2. Measures

We used three kinds of measures, including measures of the objective environment, beliefs about the environment, and weighting strategy. Fig. 1 describes the objective environment in terms of the accuracy ratio A and the bracketing rate. In our analyses we will use a variation of the accuracy ratio: let \( A' = \) the ratio of the MAD of Judge 2 (advisor) to the MAD of Judge 1 (self). This measure differs from A in that the MAD score of Judge 2 is always in the numerator, whereas with A the higher MAD score is always in the numerator. The higher the value of \( A' \), the more accurate Judge 1 is relative to Judge 2. We define Feedback \( A' \) as the value of \( A' \) calculated from the 3- or 10-trial sample feedback presented to the participants. Due to sampling error, Feedback \( A' \) will differ nonsystematically from the \( A' \) that defines the environment. To measure perceptions of relative accuracy, we constructed Perceived \( A' \), which is the ratio of the MAD scores that participants estimated. We also asked participants to state how confident they were that they had identified the better judge on a scale of 50%–100%. We transformed this to a 0%–100% scale, \( Conf_{1} \), which indicates their level of confidence that Judge 1 (self) was more accurate.

To examine how participants used the two initial estimates to develop a final answer, we computed the weight on the self/Judge 1 (\( w_{1} \)) for each item, with \( r = w_{1} \times j + (1 - w_{1}) \times a \), where \( r \) is the final revised estimate, \( j \) is the estimate of the participant (Judge 1), and \( a \) is the estimate of the advisor (Judge 2). Rearranging the terms, \( w_{1} = (r - a)/(j - a) \). Note that when revising one’s own opinion, \( w_{1} = 1 \), 0, and 0.5 correspond to choosing the self, choosing the advisor, and averaging, respectively. The distribution of \( w_{1} \) across items can be very revealing about psychological processes and has implications for accuracy, which is why we favor it over regression weights. The measure has limitations, however, which require two decisions prior to analysis (Bonaccio & Dalal, 2006; Soll & Larrick, 2009). First, the set of possible values of \( w_{1} \) may be limited by the difference between the estimates, especially if participants tend to round their final estimates. For example, if the initial estimates of APPG are 70 and 71, then rounding to the nearest whole number would permit \( w_{1} \) values of only 0 and 1. For the purpose of examining the weights, we omitted observations with differences less than two. This reflects a balance between allowing for a range of possible weights and needing a sufficient number of observations to draw conclusions. Overall, the estimates from the two judges were identical for 5.1% of trials, in which case it is not possible to compute a weight. An additional 11.5% of observations were omitted from the weight analysis due to small differences.\(^1\) Second, the final estimate occasionally falls outside the range of the

\(^1\)These observations were only omitted for the weight analysis; there were no omissions in the analysis of accuracy. We also repeated all of our analyses by including all cases where the difference between the estimates was non-zero and found very similar patterns of means and identical patterns of significance.
Table 3
Mean statistics by task and environment.

<table>
<thead>
<tr>
<th>Task/Environment</th>
<th>Environmental statistics</th>
<th>Perceptions of the environment</th>
<th>Weighting statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Target Feedback $A'$</td>
<td>Target Feedback $Br$</td>
<td>$Conf_1$ $A'$ $Br$</td>
</tr>
<tr>
<td>Revise with cues</td>
<td>Self better</td>
<td>1.70</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Equal accuracy</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Advisor better</td>
<td>0.59</td>
<td>0.30</td>
</tr>
<tr>
<td>Revise without cues</td>
<td>Self better</td>
<td>1.70</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Equal accuracy</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Advisor better</td>
<td>0.59</td>
<td>0.30</td>
</tr>
<tr>
<td>Combine others</td>
<td>Judge 1 better</td>
<td>1.70</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Equal accuracy</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Judge 2 better</td>
<td>0.59</td>
<td>0.30</td>
</tr>
</tbody>
</table>

initial estimates, leading to values of $w_1$ outside the 0–1 interval. This happened in only about 4% of cases in the two revision conditions, but in over 20% of the judgments when combining others’ opinions. For analyses involving mean weights, we truncated $w_1$ to the nearer of 0 or 1.

4.3. Results

4.3.1. Manipulation checks

Statistics relating to the environment, perceptions, and weighting are presented in Table 3. Recall that the estimates of Judge 2 (advisor) were simulated to match the desired environmental characteristics. The first four columns show the targeted environmental values, along with the simulated feedback for the accuracy ratio $A'$ and the bracketing rate $Br$. The average feedback for both variables closely matched the environments from which they were sampled.

4.3.2. Weighting strategies

The distribution of $w_1$ by condition is shown in Fig. 2, which collapses across the two levels of sampling feedback. We first note that the distributions for revising with and without cues appear to be very similar across environments. When the self is better, the participants tended to ignore the advice, with $w_1 \geq 1$ for over half of the trials (see the last column of Table 3 for the exact percentages). The frequency of choosing the self declines as the environment shifts toward favoring the advisor (moving toward the columns on the right in Fig. 2). Second, there is an asymmetry between self-favored and advisor-favored environments. The participants are egocentric in the sense that they ignore the advisor when they themselves are better, but they do not ignore the self when the advisor is—they tend to average instead. This asymmetry may reflect differences in how people interpret feedback depending on whom the feedback favors, or differences in weighting strategies, holding constant beliefs about relative ability. We will conduct an analysis in order to distinguish between these potential mechanisms later. Finally, the participants who combined others’ opinions differed in two respects from the participants in the revision conditions: they were more responsive to the environment, and they went out of range much more frequently. The frequency of out of range estimates is surprising, and we will address possible reasons for this in the discussion.

4.3.3. Revising versus combining others

We next compared revising without cues with combining others’ opinions. The design is a 2 (task) × 3 (environment) × 2 (sample size) mixed model,
with repeated measures on task (since the participants in the two tasks were yoked).

To examine the relative influence, we used the mean weight on the self, $w_1$. Although this measure masks item-level behavior such as alternating between choosing and averaging (Soll & Larrick, 2009), it does serve as an overall measure of how much participants relied on the self (Judge 1).

There was a main effect of task, such that participants placed more weight on a judge when they themselves were that judge, $F(1, 60) = 73.09, p < 0.001, \eta^2_p = 0.55$. There was also a main effect of environment, such that participants placed more weight on the more accurate judge, $F(2, 60) = 15.75, p < 0.001, \eta^2_p = 0.34$. There was no effect of sample size, and there were no significant interactions. The main effects can be seen in Table 3. In each environment, $w_1$ is about 0.2–0.3 points higher when the participants revised without cues compared to when they combined others’ opinions (e.g., in the self/Judge 1 better environment, $w_1 = 0.87$ when revising without cues, vs. 0.68 when combining others’ opinions). Moreover, moving across two adjacent environments and holding the task constant resulted in a 0.1–0.2 point change in $w_1$, such that the participants shifted weight toward the more accurate judge. This finding is supported by a test of a linear trend across the three environments within each task (revise-without-cues: $F(1, 60) = 10.59, p = 0.002, \eta^2_p = 0.15$; combine-others: $F(1, 60) = 36.17, p < 0.001, \eta^2_p = 0.37$).

One reason that participants placed more weight on their own opinion could be that they perceived the environment as being more favorable to Judge 1 (i.e., the self) in the revision task, even though the objective information in the two tasks was identical. To examine this we analyzed perceptions of the environment, including Perceived $A'$ (logged to account for skewness), confidence that the self/Judge 1 is more

Fig. 2. The distribution of $w_1$ (weight on the self/Judge 1) by condition. The tasks are depicted in the rows and the environments in the columns. The histograms collapse across levels of feedback.
accurate (Conf$^1$), and the perceived bracketing rate Br. For Perceived A’ there were main effects of task ($F(1, 60) = 5.49, p = 0.022, \eta^2_P = 0.08$) and environment ($F(2, 60) = 14.92, p < 0.001, \eta^2_P = 0.33$); all other effects were non-significant (all values of $F < 1$). As is shown in Table 3, Perceived A’ tracked the environment—on average, the participants perceived the better judge as more accurate, even when the advisor was better. However, the reversers perceived higher values of A’ than did the combiners, which indicates an egocentric bias. A similar pattern emerged for confidence, for which there were main effects of task ($F(1, 60) = 6.54, p = 0.013, \eta^2_P = 0.10$) and environment ($F(2, 60) = 12.54, p < 0.001, \eta^2_P = 0.29$), qualified by a task $\times$ environment interaction ($F(2, 60) = 2.71, p = 0.074, \eta^2_P = 0.08$). The effects involving sample size were non-significant (all values of $F < 1.2$).

The pattern underlying these results is clear from Table 3. For both Perceived A’ and Conf$^1$ the participants exhibited a biased perception of the self when they revised without cues compared to a neutral condition in which the self was not involved. In the case of confidence, this difference is increasingly pronounced as the environment shifts toward favoring the advisor. An important question is whether these biases can account for the large gap in $\bar{w}_T$ observed across the two tasks. We address this question in a supplementary analysis below.

We also examined perceptions of bracketing. Bracketing is an important element of the PAR model, and if the participants perceived different bracketing rates in the two tasks, that could potentially explain differences in weighting (prescriptively, a higher bracketing rate would justify weights closer to 50/50). As expected, there were no main or interactive effects of task on the perceived bracketing rates (all values of $F < 1$). There was an effect of sample size ($F(1, 60) = 7.88, p = 0.007, \eta^2_P = 0.12$), as well as an interaction between the environment and sample size ($F(2, 60) = 3.54, p = 0.035, \eta^2_P = 0.11$). As is shown in Table 3, the participants appear to have picked up on the higher bracketing rate in the equal accuracy environment, although the main effect of environment fell short of significance ($F(2, 60) = 2.11, p = 0.131, \eta^2_P = 0.07$). The pattern of the interaction is difficult to interpret and is immaterial to our main conclusions in any case, so we will not elaborate on it here.\footnote{The mean perceived bracketing rates for the small sample condition were 41%, 56%, and 53% in the self/Judge 1 better, equal, and advisor/Judge 2 better environments, respectively. The corresponding respective means in the large sample condition were 44%, 45%, and 33%.

### 4.3.4. Revising with and without cues

Since these conditions were not yoked, we compared them with a 2 (task) $\times$ 3 (environment) $\times$ 2 (sample size) between-subjects design. The participants who revised without cues weighted their own opinions somewhat higher ($M_{without-cues} = 0.76$ vs. $M_{with-cues} = 0.71$), although this difference fell short of significance, $F(1, 120) = 2.51, p = 0.116, \eta^2_P = 0.02$. There was a main effect of environment on the mean weight: as is shown in Table 3, the participants placed more weight on their own opinions the higher the environmental $A'$, $F(2, 120) = 18.51, p < 0.001, \eta^2_P = 0.24$. This effect was qualified by an environment $\times$ sample size interaction, $F(2, 120) = 3.63, p = 0.029, \eta^2_P = 0.06$. The mean weights on the self were 0.87 (Self Better environment), 0.76 (Equal Accuracy), and 0.52 (Advisor Better) when 10 feedback trials were presented, compared to 0.85, 0.74, and 0.69, respectively, with 3 feedback trials. Follow-up tests of the simple effects showed that sample size had no effect in the Self Better and Equal Accuracy environments ($F < 1$), but did matter when the advisor was better, $F(1, 120) = 8.89, p = 0.003, \eta^2_P = 0.07$.

We also examined perceptions of the environment. Since both conditions involved revision, we expected to find environment but not task differences. In the case of Perceived A’ there was only a main effect of environment, $F(2, 120) = 21.87, p < 0.001, \eta^2_P = 0.27$. However, for confidence in the self we found main effects of both environment ($F(2, 120) = 17.14, p < 0.001, \eta^2_P = 0.22$) and task ($F(1, 120) = 5.29, p = 0.023, \eta^2_P = 0.04$). Participants were more confident that they were the better judge when they were favored by the environment and when they did not have access to the cues during the revision stage ($M = 59.8$ vs. 49.5).

To summarize this analysis, two additional findings have emerged. First, as another example of an egocentric bias, sample size had asymmetric effects on
the weighting, depending on whom the feedback favored. When the feedback favored the self, the participants ignored the advisor equally, regardless of whether they received 3 or 10 feedback trials. However, when the feedback favored the advisor, the participants were more convinced when they had more data on past accuracy. Second, the participants were more self-confident when they lacked cues at the revision stage, which may have accounted for the slightly lower level of advice taking in this condition.

4.3.5. Can biased perceptions explain the egocentric discounting of advice?

We showed earlier that the participants perceived a given judge as relatively more accurate when they themselves were that judge. We now examine the extent to which such egocentric perceptions can explain a high weighting of the self. This analysis includes only the revise-without-cues and combine-others conditions, because we wanted to isolate the pure effect of revision that is not confounded with other factors (such as the presence of cues in the revise-with-cues condition).

The fundamental question driving the analysis is the following: what would the weight on the self ($w_{1}$) have looked like in the revision task had perceptions of the environment and confidence been similar to those in the combine-others task? The answer to this question will reveal the extent to which biased perceptions of accuracy are the source of egocentric advice discounting. We answered the question in three steps. First, we regressed the Perceived $A'$ and $Conf_{1}$ on Feedback $A'$, task (dummy coded: $1 = \text{revise}$, $0 = \text{combine-others}$), and their interaction (Feedback $A'$ and Perceived $A'$ were logged, and a log-odds transformation was applied to $Conf_{1}$ to ensure predicted values in the range 0–100). The results are shown in Equations 1 and 2 of Table 4 and are plotted in Fig. 3. The plots show that the participants inflated their self-perceptions and confidence more when the feedback favored the advisor, relative to a benchmark in which the same information was learned about Judge 1 in the combine-others task. Second, we regressed the weight on the self/Judge 1 on Feedback $A'$, Perceived $A'$, $Conf_{1}$, and task (see Table 4, Equations 3–5). The best fit is provided by Equation 5, which shows that perceived relative accuracy, confidence, and task are sufficient to predict weight. The weight does depend on the objective measure Feedback $A'$ (Equation 3), but, unsurprisingly, this is mediated by the subjective variables (i.e., Feedback $A'$ drops out once perceptions are included in the model). Third, we used the predicted values of Perceived $A'$ and $Conf_{1}$ from Equations 1 and 2 for different levels of Feedback $A'$ and substituted these into Equation 5. This allows us to plot estimated curves of the weight on the self/Judge 1 as a function of objective feedback, task, and the presence or absence of biased perceptions, as shown in Fig. 4.

In Fig. 4, the combine-others curve sets task equal to zero in Equations 1, 2, and 5, and the revise biased curve sets task equal to one for all three equations. These two curves represent the predicted weight for each task as a function of the objective feedback. The gap between the curves confirms that people put more weight on an opinion if it is their own, holding feedback constant. To estimate the revise unbiased curve, we set task equal to zero in Equations 1 and 2 and equal to one in Equation 5. The resulting curve estimates how participants in the revision task would have weighted the opinions had they not had inflated perceptions and confidence. Even with unbiased perceptions and holding objective feedback constant, people still weight an opinion more if it is their own. Nevertheless, biased perceptions do account for a portion of advice discounting, though mainly when the feedback favors the advisor, as reflected in the gap between the two revision curves on the left side of the graph.

4.3.6. Accuracy

To examine performance, we computed the percentage improvement over choosing a judge at random for intuitive aggregation (i.e., the participant’s final judgment) and three benchmark strategies: averaging, choosing based on performance in the 3 or 10 feedback trials, and perfect choosing (choosing the judge who performed better on the 20 evaluation trials). Although perfect choosing is unattainable in

---

3 We also examined alternative models that substituted for Perceived $A'$ its constituent parts, the logged participant estimates of the MADs of the two judges. Because the MADs of the self and the advisor received similar weights in these alternative models, models that employ Perceived $A'$ are not only simpler, but also similarly predictive.
Fig. 3. Plots of Perceived $A'$ and $Conf_1$ as a function of Feedback $A'$ and the task, based on the regression models in Table 4. Both Perceived $A'$ and Feedback $A'$ are shown on a log scale, which ensures that points which are equidistant from the vertical axis reflect equal accuracy ratios. For example, when $A' = 0.5$, Judge 1’s MAD is twice that of Judge 2, and when $A' = 2$, Judge 2’s MAD is twice that of Judge 1.

Table 4
Regression models.

<table>
<thead>
<tr>
<th>Equation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td>ln (Perceived $A'$)</td>
<td>ln ($\frac{Conf_1}{100-Conf_1}$)</td>
<td>$\ln^{1}$</td>
<td>$\ln^{1}$</td>
<td>$\ln^{1}$</td>
</tr>
<tr>
<td>Intercept</td>
<td>$-0.072^*$</td>
<td>$-0.086$</td>
<td>$0.515^{****}$</td>
<td>$0.430^{****}$</td>
<td>$0.421^{****}$</td>
</tr>
<tr>
<td>ln (Feedback $A'$)</td>
<td>$0.760^{****}$</td>
<td>$3.468^{****}$</td>
<td>$0.260^{****}$</td>
<td>$0.035$</td>
<td>$0.170^{**}$</td>
</tr>
<tr>
<td>ln (Perceived $A'$)</td>
<td>$0.174^{****}$</td>
<td>$0.248^{****}$</td>
<td>$0.202^{****}$</td>
<td>$0.174^{****}$</td>
<td>$0.198^{****}$</td>
</tr>
<tr>
<td>$Conf_1$</td>
<td>$0.150^{**}$</td>
<td>$0.514^*$</td>
<td>$0.248^{****}$</td>
<td>$0.202^{****}$</td>
<td>$0.198^{****}$</td>
</tr>
<tr>
<td>Task</td>
<td>$0.150^{**}$</td>
<td>$0.514^*$</td>
<td>$0.248^{****}$</td>
<td>$0.202^{****}$</td>
<td>$0.198^{****}$</td>
</tr>
<tr>
<td>ln (Feedback $A'$) $\times$ task</td>
<td>$-0.232^{***}$</td>
<td>$-1.684^{****}$</td>
<td>$-0.118^{**}$</td>
<td>$-0.034$</td>
<td>$0.016$</td>
</tr>
<tr>
<td>ln (Perceived $A'$) $\times$ task</td>
<td>$0.634$</td>
<td>$0.534$</td>
<td>$0.505$</td>
<td>$0.627$</td>
<td>$0.629$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.634$</td>
<td>$0.534$</td>
<td>$0.505$</td>
<td>$0.627$</td>
<td>$0.629$</td>
</tr>
</tbody>
</table>

$n = 132.$

$^{****} p < 0.001.$

$^{***} p < 0.01.$

$^{**} p < 0.05.$

$^* p < 0.10.$

the absence of perfectly diagnostic cues to expertise, it is a useful benchmark nevertheless. The results are shown in Table 5, which provides statistical tests comparing intuitive aggregation to each of the three benchmark strategies (the comparisons among benchmarks are either logically necessary or implied by the PAR model). As the results are fairly detailed, we will focus here on several key aspects.

First, whether averaging or perfect choosing performs better depends on the environment. This is a key insight from the PAR model (Soll & Larrick, 2009)—averaging may still be the better strategy, even when the more accurate judge is known with certainty. In the current study, averaging outperforms perfect choosing in the Equal Accuracy environment. The other two environments had sufficiently large differences in expertise (and lower bracketing rates) that choosing was the better strategy, provided that the better judge could be identified with certainty.

Second, the appropriate response to feedback depends on both the amount of feedback and the generating environment. With just three trials, the feedback occasionally favored the less accurate judge, and thus choosing based on feedback fell short of per-
fect choosing across environments and tasks. Moreover, averaging substantially outperformed choosing according to feedback when the judges were equally accurate, and achieved similar performance elsewhere. Across environments, when there were three feedback trials, averaging performed either about as well as or much better than choosing based on the feedback. In contrast, with ten trials the feedback discriminates among judges’ skill levels more accurately, and choosing based on the feedback approaches the performance of perfect choosing. In this case, averaging continues to perform best in the Equal Accuracy environment, but choosing based on feedback has the advantage in the large-difference environments (i.e., Self Better, Advisor Better). In some environments, therefore, it is realistic to choose and outperform the simple average, provided that there is a reasonably large sample of evidence on performance (Fischer & Harvey, 1999).

Third, the performance of the participants’ intuitive strategies depended very much on the task, environment, and amount of feedback. In the Equal Accuracy environment, the participants who combined others’ opinions outperformed those who revised, although both groups fell far short of pure averaging. To their credit, the participants who combined others’ opinions often averaged in the Equal Accuracy environment (see the middle column of Fig. 2); however, they also went out of range more frequently, which harmed their accuracy. The results are more complex in the large-difference environments. Those who combined others’ opinions either matched averaging (3 feedback trials) or slightly surpassed it (10 trials). In contrast, the performance of the revisers interacted with both the environment and the sample size. Not surprisingly, the revisers performed particularly well in the Self Better environment, regardless of the sample size. As a general rule, intuitive weighting strategies are biased toward the self, so people will naturally do well when high self-weighting happens to be appropriate. In the Advisor Better environment, the revisers underperformed averaging, but only when the sample size was small. In many situations, the participants who combined others’ opinions outperformed the revisers because they were more willing to average and were less biased in their perceptions. However, the revisers performed well when they were substantially more accurate than their advisors, because in this case egocentric discounting of advice worked in their favor. Since people tend to weight their own opinions highly regardless of the environment, the strong performance of the revisers in this environment is due more to luck than to skill—these participants were fortunate enough to be assigned to an environment in which their egocentric strategy happened to be ideal.

Finally, it is notable, although not unexpected, that the participants tended to perform better when they had access to larger samples of feedback. We had conjectured that small samples might lead to more extreme weighting of the judges, and consequently lower accuracy, assuming that the observed feedback was transformed directly into weights without correcting for sample size (cf. Griffin & Tversky, 1992). As expected, small samples did have more extreme accuracy ratios (the value of $A$—larger $MAD$ over smaller—observed in the feedback). In the Equal Accuracy environment, $A = 1.72$ with 3 feedback trials, compared to $A = 1.37$ with 10 trials (the corresponding values for the large-difference environments were 2.16 and 1.72). However, additional analyses (omitted for brevity) indicated that the extremity of weights (i.e., $w_1$ closer to 0 or 1) was similar in the two sample size conditions.
5. Discussion

The experiment has three main findings. First, people approach the revision and combining others tasks differently. When revising, people tend either to be unmoved by advice or to average. In contrast, people are more likely to average when they combine the opinions of others. Together, these differences constitute the egocentric discounting of advice: people tend to place more weight on an opinion when it is their own, holding constant the available objective information about the accuracy of the opinions. Second, people are egocentrically biased in forming their beliefs about relative accuracy. Compared to participants asked to combine the opinions of others, those who revised tended to inflate their perceptions of their own relative expertise and confidence. These biases were most prominent when feedback favored the advisor. Third, biased perceptions about relative expertise and confidence can account for only a small portion of the egocentric discounting of advice. There is a task main effect such that people place more weight on their own opinions even when beliefs and confidence are held constant. We next elaborate on these points, and also discuss prescriptions based on our findings.

5.1. Explaining egocentric discounting of advice

The participants formed self-serving beliefs, especially when confronted with a negative social comparison. They did not necessarily elevate themselves above others—when the objective data favored the advisor, the participants still conceded that the advisor was more accurate than they were. They did, however, suppress the magnitude of their inferiority. This finding is consonant with Kunda’s (1990) view that motivated reasoning is constrained by one’s ability to stretch or interpret the available evidence in a way that seems reasonable or justifiable. For example, one might be able to attribute a single inaccurate judgment to a lapse of attention or to bad luck, but it would be difficult to do this for the feedback as a whole.

Despite the fact that beliefs are biased in favor of the self, this accounts for only a small proportion of the egocentric discounting of others’ opinions. As can be seen in Fig. 4, people weight an opinion about 20 percentage points higher (e.g., a weight of 0.5 on the self as opposed to 0.3) when it is their own, controlling for beliefs and confidence. Self-serving beliefs can add up to 10 additional percentage points in environments in which the advisor performs substantially better on objective feedback. Our results confirm past research showing biased perceptions of one’s own ability (Burson et al., 2006; Krueger & Mueller, 2002), but these effects are apparently not responsible for the bulk of the egocentric discounting of advice.

Similarly, differential access to reasons also falls short as an explanation. We observed substantial advice discounting in the revise-without-cues condition, in which participants did not have access to the original cues when they revised. The mean weight on the self in this condition was 0.71, which is similar to the findings of past studies in which participants had access to reasons (Harvey & Fischer, 1997; Yaniv & Kleinberger, 2000). Clearly, access to reasons during the revision stage is not a necessary condition for discounting advice egocentrically. Another possibility is that having access to reasons during initial judgments enhances the perception of one’s own ability, perhaps as a result of cogency or fluency of processing (Alter & Oppenheimer, 2009). In this way, access to reasons can have an indirect effect on opinion revision even when cues are absent in the revision stage, with perceptions of relative ability and confidence serving as mediators. However, we have already found that these subjective variables cannot account for the large task main effect. This result would seem to preclude an explanation based on differentially weighting estimates as a direct function of the supporting evidence (which includes both reasons for the judgment and beliefs about relative skills).

Although differential access to reasons cannot fully explain the egocentric discounting of others’ opinions, it may still contribute to it. In order to begin to explore this possibility, we included a condition in which participants had access to the cues when they revised. We had anticipated that people would weight their own opinions more when they could revisit the rationale for their judgment during the revision stage. In fact, the results suggest that the opposite may be true—if anything, participants weighted their own opinions more when they lacked the cues. It may be that the processing of reasons differs depending on whether they are internally generated
or externally provided.\textsuperscript{4} The participants in our study received feedback and could see how the cues and the criterion were imperfectly related. The feedback for internally generated reasons is likely to be more ambiguous and leave more room for ego-enhancing interpretation. One path for future research, therefore, would be to investigate whether differential access to the private thoughts of the judges contributes to advice discounting.

Given that the two leading potential explanations cannot account for most of discounting, what can? We suggest three possibilities, although more could undoubtedly be added to the list. First, statistical intuitions are sometimes malleable and can be shaped to support a preferred course of action (Sanitioso & Kunda, 1991). If people have a preference for their own opinions, then they might maximize the weight on their own opinions by selectively choosing between conflicting intuitions about the value of averaging (Larrick & Soll, 2006; Soll, 1999). For example, when the self is perceived as being more accurate, people may reason that “compromise leads to mediocrity”, but when the advisor appears to be more expert, they may think that “the truth lies in the middle”.

Second, people may discount advice because they anticipate unfavorable comparisons with how accurate they would have been had they kept their initial answers. One factor that plays a part in this is whether people expect to receive feedback on both their initial answers. How this affects weighting policies depends on which level of accuracy serves as a reference point. Suppose that people code a change of opinion as action and ignoring advice as inaction, in which case the accuracy of one’s initial judgements is likely to be perceived as the reference point. In that case, changing one’s answers and reducing accuracy would be evaluated by a loss-averse individual as worse than retaining one’s initial answers and failing to improve (Baron & Ritov, 1994). This asymmetry in how changes in accuracy are evaluated would lead to the underweighting of advice.

Third, people may simply like their own opinions better, independent of their perceived accuracy—a mere ownership effect (Beggar, 1992; Heider, 1958; Nuttin, 1987). People may see their opinions as possessions, and hold them in higher esteem by virtue of this association (Abelson & Prentice, 1989). If this is the case, then ceding ground to advice may be perceived as devaluing one’s own opinion, and by extension devaluing oneself. This could explain why people prefer their own opinions, and lead to the motivated statistical reasoning we suggested earlier. Alternatively, mere ownership may be a sufficient explanation for the egocentric discounting of advice: people may be willing to accept lower accuracy in exchange for keeping the opinions that belong to them, as a way of preserving self-esteem. If taking advice is ego-threatening, one way of countering the threat would be to affirm an individual in a different domain (Steele, 1988). Consistent with this, Soll and Larrick (2009) found that participants chose the advisor more frequently when they answered questions from one’s own area of expertise, people may feel less threatened by fully accepting advice in a different area.

5.2. Out of range estimates

The final estimates were outside the range of the initial estimates in about 5% of judgment trials for the revision tasks (both with and without cues), a result which is consistent with past research. In contrast, the participants who combined others’ opinions went out of range on about 25% of judgments. Out of range estimates have previously been reported in studies of group judgment (Sniezek & Henry, 1989), jury awards (Schkade, Sunstein, & Kahneman, 2002), and multiple cue learning tasks (Lees & Triggs, 1997). There are, of course, good reasons for going out of range. For example, new information may come to light that leads one to adjust away from rather than toward an advisor. However, this cannot explain our data, as out of range estimates were far more prevalent in the combine-others task, where the participants did not have any additional information. Another

\textsuperscript{4} We thank an anonymous reviewer for suggesting this possibility.
valid reason would be statistical—either to regress toward the mean outcome or to correct for bias. People often reproduce the variance of the outcome in their forecasts (Harvey, 1995), and if the participants suspected this of the judges then ideally they would have regressed, occasionally going out of range to do so. The data do not support this interpretation either: a follow-up analysis showed that out of range estimates were just as likely to be away from the mean as toward it. What is predictive is the range of the two estimates—out of range estimates were more common for smaller ranges.

One possible explanation for out of range estimates is probability matching (Shanks, Tunney, & McCarthy, 2002). The participants knew that occasionally the two estimates would not bracket the truth, and they may have tried to guess on individual trials when that might happen. Of course, probability matching is a losing strategy because it amounts to chasing noise; the participants would have done better with a consistent policy of choosing or averaging, as appropriate to the environment. Given that out of range estimates can have a substantial effect on accuracy, future research should investigate when they are most likely to occur.

5.3. Prescriptions

Our recommendations to practitioners depend on how accurately one can know the environment. When the environment is opaque, as in the small-sample conditions, then it is best to simply average. As is shown in Table 5, even when there are large differences in the expertise, small samples of performance feedback cannot reliably distinguish between these levels. Thus the advantage of choosing based on feedback is small and is more than outweighed by the benefits accrued when the judges are similar in ability. It is interesting to note that with small samples, the participants in both the revision and combine-others tasks could have improved by averaging. There was one exception: revisers who were randomly assigned to the Self Better environment, and who therefore had advisors who were 70% less accurate than themselves; it is not surprising that they outperformed averaging. Their strong performance is more than outweighed, however, by the relatively poor accuracy achieved by the participants in the other two environments. The participants in the Self Better environment were lucky that their egocentric strategy, which they applied universally, happened to match well with the environment to which they were randomly assigned.

With a large sample, one can begin to get a sense of the environment, and a choosing strategy may be appropriate. Table 5 shows that choosing based on feedback approached the performance of perfect choosing in the large discrepancy environments. However, the practitioner does not typically know the environment, but rather must make a guess about it based on the feedback. If the practitioner works in environments which are characterized by large differences in expertise (e.g., medicine, academia), then a sample of 10 may be sufficient to conclude that one judge is substantially better than the other. However, the same inference cannot be made in environments characterized by a low dispersion in expertise (e.g., possibly mutual fund managers), in which case sample differences, even in large samples, are likely to be due to luck. Because our experiment did not systematically investigate the diagnosticity of feedback across a wide range of dispersion levels, we cannot derive any general advice from it. We can, however, offer two more modest prescriptions. First, one should ask whether the cues to expertise are really as diagnostic as they may seem. People often rely on cues such as confidence and talkativeness, which are known to be poorly correlated with expertise (Klayman, Soll, González-Vallejo, & Barlas, 1999; Littlepage, Schmidt, Whisler, & Frost, 1995). Unless valid cues clearly point to one judge being substantially better than the other, averaging is probably the better strategy. Second, in revising one’s own opinion there may be situations in which the advisor is clearly much better than oneself, in which case there may be substantial gains from choosing the advisor and putting little weight on one’s own opinion. This was the case with 10 sample trials—the participants improved on a random judge by 9.4% when the advisor was better, but they could have improved by 25.4% by choosing according to the feedback. Of course, this only makes sense if the evidence clearly favors the advisor and the environment is one in which large differences in expertise are likely.

This discussion underscores the way in which the effectiveness of aggregation strategies is contingent on the environment. Potential avenues for future research include developing methods for identifying the environment and developing aggregation strategies that are robust across environments. Of course, there
Table 5
Mean percentage improvement over choosing a judge at random.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Task/environment</th>
<th>n</th>
<th>Intuitive aggregation</th>
<th>Benchmark strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Averaging</td>
<td>Choosing (feedback)</td>
</tr>
<tr>
<td>3 trials</td>
<td>Revise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self better</td>
<td>22</td>
<td>23.6</td>
<td>11.4**</td>
<td>14.5</td>
</tr>
<tr>
<td>Equal accuracy</td>
<td>22</td>
<td>15.1</td>
<td>26.9**</td>
<td>−2.5**</td>
</tr>
<tr>
<td>Advisor better</td>
<td>22</td>
<td>1.0</td>
<td>12.3**</td>
<td>13.8*</td>
</tr>
<tr>
<td>Combine others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 1 or 2 better</td>
<td>22</td>
<td>9.6</td>
<td>10.1</td>
<td>9.5</td>
</tr>
<tr>
<td>Equal accuracy</td>
<td>11</td>
<td>16.1</td>
<td>27.3**</td>
<td>−1.0**</td>
</tr>
<tr>
<td>10 trials</td>
<td>Revise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self better</td>
<td>22</td>
<td>24.0</td>
<td>8.8**</td>
<td>24.5</td>
</tr>
<tr>
<td>Equal accuracy</td>
<td>22</td>
<td>13.1</td>
<td>29.4**</td>
<td>−4.8**</td>
</tr>
<tr>
<td>Advisor better</td>
<td>22</td>
<td>9.4</td>
<td>9.0</td>
<td>25.4**</td>
</tr>
<tr>
<td>Combine others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 1 or 2 better</td>
<td>22</td>
<td>13.1</td>
<td>8.0</td>
<td>23.9*</td>
</tr>
<tr>
<td>Equal accuracy</td>
<td>11</td>
<td>20.4</td>
<td>29.5*</td>
<td>−5.8*</td>
</tr>
</tbody>
</table>

The marked entries indicate that the benchmark strategy differed significantly from intuitive aggregation with a paired t-test (2-tailed).

** p < 0.01.
* p < 0.05.

are Bayesian approaches to aggregation that address these issues (Clemen & Winkler, 1999), but we believe that practitioners would benefit from simple strategies that teach them how to aggregate on an everyday basis. For example, one such strategy when there are many forecasters and only a short history of past judgments is to average the opinions of the top five historical performers (Mannes, Soll, & Larrick, unpublished). In contrast to the strategies of averaging all forecasters and choosing just one, the top-five strategy is robust, in the sense that it consistently performs well across a wide range of environments.

5.4. Conclusion

People persist in their opinions, even in the face of good advice. Only rarely do people move more than halfway toward advice, even when the advisor is proven to be substantially more accurate than oneself. Neither differential access to reasons nor egocentric beliefs can fully account for the robust tendency to overweight one’s own opinions. Echoing past research, we advise practitioners to think twice before choosing a single expert: unless there are large differences in expertise and the expert can be identified reliably, it is very hard to beat a simple average.

Acknowledgements

We thank Rick Larrick, Stefan Herzog, Julia Minson, Don Moore, and Kelly See for discussions, comments, and suggestions that were helpful to us in writing this article.

Appendix

For each participant who completed the initial phase of the revision task, we simulated an advisor using the participant’s mean absolute deviation, $MAD_1$, over the 30 judgment trials, as an input. The advisor’s $MAD$ score was targeted to be $A' \times MAD_1$, with estimates that bracket the truth at the desired rate. In this study, $A'$ was fixed at 1 (Equal Accuracy), 1.7 (Self Better), and 1/1.7 = 0.59 (Advisor Better). For each judgment trial, let $Y_1$ be the participant’s estimate, $T$ the criterion, and $E_1$ the signed error, defined as $Y_1 - T$. Let $Z_{e1}$ be the participant’s standardized error, computed using the mean and population standard deviation from the set of 30 judgment trials. We used the following formula to compute the advisor’s standardized error on a given question:

$$Z_{e2} = r Z_{e1} + \sqrt{1 - r^2} Z,$$
where \( r \) is the correlation needed to generate the desired bracketing rate and \( Z \) is a standard normal deviate which is randomly generated for each question. We used \( r = 0 \) and \( r = 0.65 \) to obtain bracketing rates of 50\% and 30\%, respectively. To transform the advisor’s standardized errors to raw errors, we used the formula

\[
E_2 = A' \times MAD_1 \times Z_{e2}\pi/2.
\]

Assuming that \( E_2 \) is approximately normal, the mean absolute deviation of \( E_2 \) is \( MAD_2 = A' \times MAD_1 \), as desired. The advice for a given question is \( T + E_2 \).

References


