Journal of Personality and Social Psychology
2014, Vol. 107, No. 2, 276–299

0022-3514/14/$12.00 http://dx.doi.org/10.1037/a0036677

The Wisdom of Select Crowds

Albert E. Mannes
Philadelphia, Pennsylvania

Jack B. Soll and Richard P. Larrick
Duke University

Social psychologists have long recognized the power of *statisticized* groups. When individual judgments about some fact (e.g., the unemployment rate for next quarter) are averaged together, the average opinion is typically more accurate than most of the individual estimates, a pattern often referred to as the *wisdom of crowds.* The accuracy of averaging also often exceeds that of the individual perceived as most knowledgeable in the group. However, neither averaging nor relying on a single judge is a robust strategy; each performs well in some settings and poorly in others. As an alternative, we introduce the *select-crowd* strategy, which ranks judges based on a cue to ability (e.g., the accuracy of several recent judgments) and averages the opinions of the top judges, such as the top 5. Through both simulation and an analysis of 90 archival data sets, we show that select crowds of 5 knowledgeable judges yield very accurate judgments across a wide range of possible settings—the strategy is both accurate and robust. Following this, we examine how people prefer to use information from a crowd. Previous research suggests that people are distrustful of crowds and of mechanical processes such as averaging. We show in 3 experiments that, as expected, people are drawn to experts and dislike crowd averages—but, critically, they view the select-crowd strategy favorably and are willing to use it. The select-crowd strategy is thus accurate, robust, and appealing as a mechanism for helping individuals tap collective wisdom.

Keywords: judgment, decision making, aggregation, expertise, groups

Each year, *The Wall Street Journal* conducts a forecasting competition for economists. There are typically around 50 participants representing some of the nation's most elite academic, government, and business institutions. The task is to predict a host of economic variables over the coming year, such as the rates of unemployment and economic growth in the United States. A feature story is later published that celebrates the foresight of the economist who prognosticated the most accurately. Of course, many important choices are influenced by economic forecasts, from companies' hiring decisions to the Federal Reserve Board's position on interest rates. This raises two interesting questions for social scientists—whose opinion do people follow when making these decisions, and whose should they follow if they desire to maximize their accuracy?

We begin this article with the prescriptive question inspired by *The Wall Street Journal*'s contest: How should someone confronted with a set of diverse opinions use them in order to make the best judgment possible? Past social psychological research offers two solutions to this problem. One strategy is to seek out the most knowledgeable person in a group and rely on his or her judgment. This is commonly referred to as a *best-member* strategy (Yetton & Bottger, 1982), and considerable research has focused on ways to improve groups' ability to identify and leverage their expertise (e.g., Bonner, 2004; Hackman, 1987; Henry, 1995; Libby, Trotman, & Zimmer, 1987; Steiner, 1972). Socrates himself was perhaps the strategy's first advocate:

> And for this reason, as I imagine,—because a good decision is based on knowledge and not on numbers? . . . Must we not then first of all ask whether there is any one of us who has knowledge of that about which we are deliberating? If there is, let us take his advice, though he be one only, and not mind the rest; if there is not, let us seek further counsel. (Plato, 2005, p. 46)

This strategy can easily be extended beyond the group context to any situation in which a decision maker is confronted with a crowd of opinions. To judge well, the decision maker should identify the crowd's most qualified member and defer to his or her opinion.

An increasingly popular alternative strategy is to leverage collective knowledge by relying on the *wisdom of crowds* (Surowiecki, 2004). This phenomenon was famously demonstrated by Francis Galton (1907), who reported that the dressed weight of an ox on display at the local fair was only one pound more than the mean estimate of nearly 800 spectators. Since then, exploring the benefits of *statisticized groups* has been a mainstay of social psychological research (e.g., Davis, 1996; Hastie, 1986; Hinsz, 1999; Hogarth, 1978; Lorge, Fox, Davitz, & Brenner, 1958; Stroop, 1932; Wallsten, Budescu, Erev, & Diederich, 1997; Yaniv, 2004). Numerous studies have demonstrated that simple rules for aggregating judgments (such as the median or mean for numerical

Albert E. Mannes, Philadelphia, Pennsylvania; Jack B. Soll and Richard P. Larrick, Fuqua School of Business, Duke University.

This research benefited greatly from the input and feedback we received at numerous seminars and conference presentations. In particular, we wish to acknowledge David Budescu, Robin Hogarth, Cade Massey, Barbara Mellers, Philip Tetlock, and George Wu for their insightful contributions. We also thank Don Moore and his colleagues for sharing data from their studies with us.

Correspondence concerning this article should be addressed to Jack B. Soll, Fuqua School of Business, Duke University, Durham, NC 27708. E-mail: jsoll@duke.edu

judgments or majority vote for categorical ones) perform as well as or better than more complex strategies (for reviews, see Clemen, 1989; Gigone & Hastie, 1997; Hastie, 1986; Hastie & Kameda, 2005; Hill, 1982; Kerr & Tindale, 2011; Wallsten et al., 1997). For instance, the mean of a crowd's forecasts will typically prove superior in quality to the forecast of the crowd's average member.

Both strategies entail risk. On the one hand, a decision maker who listens to only one person may miss out on knowledge dispersed among many and may perform very poorly if an inferior member of the crowd is accidentally chosen. On the other hand, an average of all opinions may include some very ill-informed ones, which could substantially harm accuracy. In this article, we introduce a third approach, which we call the *select-crowd* strategy. It is implemented by averaging the opinions of a small number of individuals chosen for their ability or expertise. We demonstrate with both simulated and archival data that the strategy of relying on the single person chosen as best and the strategy of averaging the whole crowd each performs well in some judgment environments but poorly in others. Select crowds, in contrast, perform well across many judgment environments—they are often the best and rarely the worst—which makes them an appealing option when the environment is unknown.

We also pursue in this article the descriptive question: Whose opinion do people follow when they confront a crowd of opinions, and what psychological factors account for their behavior? Previous empirical research suggests that people are reluctant to use judgments averaged across a large group and endorse relying on a single judge believed to be the best (e.g., Larrick & Soll, 2006; Soll & Mannes, 2011). We show, however, that when the option to select a subset of judges from the crowd is available, people prefer to average the opinions of two or three top performers. We demonstrate this result in two experiments and discuss the reasons for it in a third. Although people on average include too few judges in their select crowds, the strategy is a psychologically attractive alternative to relying on a single expert.

We begin the article by describing the judgment problem in detail and our model for describing the characteristics of the environment. The *environment* consists of factors such as whether the members of a crowd differ greatly or very little in ability and whether they tend to make similar or different judgment errors. This is important because different strategies perform well in different environments. Next, we examine the comparative performance of three judgment strategies across a wide range of environments: choosing the one member of the crowd believed to be best (the best-member strategy), averaging the opinions of all members of the crowd (the whole-crowd strategy), and averaging the opinions of a subset of members from the crowd selected for their ability (the select-crowd strategy). The strategies are first tested in simulated environments and then with archival data. For this purpose, we employed 40 data sets from published work in psychology and 50 data sets from the *Survey of Professional Forecasters*, which is conducted by the Federal Reserve Board of Philadelphia. These analyses all point to the effectiveness of the select-crowd strategy across a wide range of environments. Finally, we report the results of three studies that describe which strategies people endorse and use. Overall, we find that select crowds are both a prescriptively wise and an intuitively appealing strategy for improving judgment.

The Judgment Problem

We examine situations in which a decision maker wishes to estimate some quantity as accurately as possible, such as the unemployment rate next quarter, the life expectancy of an ill patient, or the number of jellybeans in a jar. Accuracy is defined as closeness to the realized or true value of the criterion, as measured by absolute error (AE; the unsigned difference between the estimate and the truth). We assume the decision maker is uncertain about the correct answer and has access to the opinions of a number of judges, collectively known as the *crowd*, which he or she can consult. Membership in the crowd depends on the context, but the general notion is that a crowd consists of *all* members of a particular collective. For example, the crowd can be all students in a class, all economists participating in *The Wall Street Journal*'s forecasting competition, or all doctors at a particular hospital.

The crowd's members may vary in their expertise, which we define as the ability to make accurate estimates or forecasts. A key parameter in our model of the environment is *dispersion in expertise*, which quantifies the degree to which a crowd's members differ in their ability to estimate accurately. At one end of the continuum, all judges have identical expertise. Such a crowd could consist of either all experts or all novices. As long as the level of expertise in the crowd is constant, regardless of that level, then there is zero dispersion. For example, kindergarteners estimating the dollar value of coins in a jar may all be similarly poor at the task, whereas professional meteorologists forecasting tomorrow's temperature may all be similarly capable. Although the two cases exhibit very different levels of expertise, they are both characterized by low dispersion. At the other end of the continuum, judges vary greatly in their expertise. This may be the case, for example, in a college trivia bowl when history majors compete against physics majors on the dates of historical events. Typically, the degree of dispersion in expertise will be somewhere between these two extremes, with some judges being more capable than others but not dramatically so.

Environments with high dispersion in expertise present an opportunity to identify and capitalize on the abilities of the more capable judges (Mellers et al., 2014). A good strategy might be to adopt the opinion of the one judge *predicted* to be the best, such as the individual who has performed the best historically. According to this strategy, the decision maker ranks the crowd's members based on a cue to expertise. Although we focus in this article on historical performance as the cue, in principle the cue can be any judge-specific attribute, such as confidence, status, or credentials. Most likely the cue will be fallible, in the sense that there is a risk of choosing someone who is not the crowd's true expert. As a shorthand, we refer to this strategy as the *best-member strategy*. What we mean by this is that the decision maker relies on the opinion of the judge ranked first in the crowd by an imperfect cue. We emphasize that the best member is identified *prospectively* in our model. The chosen judge may or may not be the crowd's true expert; the chance of identifying that person depends critically on the validity of the decision maker's cue. If the cue to expertise is at all valid, the judge who ranks first on the cue is more likely than anyone else to be the individual with greatest ability. Typically, it is easier to discriminate among judges when they differ greatly in ability. Thus, cues to expertise will tend to be more valid in high-dispersion environments.

With the best-member strategy, all members of the crowd are ignored except for the one chosen based on available cues. The opposite of this strategy would be to weigh everyone's opinion equally—to average them, which is the wisdom-of-crowds approach. Averaging tends to outperform both the crowd's average member and the judgments of interacting groups arrived at through discussion and consensus. This is because interacting groups are prone to social-influence processes and judgment biases that undermine accuracy (for reviews, see Hastie, 1986; Kerr, MacCoun, & Kramer, 1996; Kerr & Tindale, 2004; Steiner, 1972). The loss associated with these factors is often quantified by comparing the accuracy of a group's consensus-based judgments with the accuracy of its corresponding statisticized judgments in which members' individual opinions are simply averaged (see, e.g., Gigone & Hastie, 1997).

For numerical judgments, the performance of a simple average depends on two factors: the mean level of expertise in the crowd and the independence of its members (i.e., whether they make similar or dissimilar judgment errors). The crowd's mean level of expertise sets a floor on the performance of averaging, so the smarter the crowd, the wiser its average judgment. Independence increases the chance that individual judgments bracket the truth (Larrick & Soll, 2006), some falling above and some below the correct answer, which allows their errors to cancel out in the aggregate. Consider, for example, two people predicting the high temperature tomorrow (in degrees Fahrenheit). Person 1 predicts 70, Person 2 predicts 66, and the actual high temperature turns out to be 78. Because Person 1's AE is eight and Person 2's AE is 12, the average AE in this group is 10, which is a measure of this group's expertise. Their average prediction is 68, the AE of which is also 10. This is not a coincidence. In this case, both people made similar errors (they both underestimated the actual temperature), so there was no bracketing of the truth. In the absence of bracketing, an average prediction will be as accurate as the average person in the group (in this scenario, 10°). The expertise of the group sets the floor on the performance of averaging.

With bracketing, positive errors by one person are offset to some degree by the negative errors of another. In the aforementioned example, if Person 1 predicts 70 (as before), Person 2 predicts 90 (instead of 66), and the actual temperature remains 78, their average AE is still 10. However, the AE of their average prediction of 80 is only two. In this case, their average prediction is more accurate than the average person in the group because their estimates bracket the truth (Person 1 underestimated the actual temperature, and Person 2 overestimated it). This illustrates the error-cancellation benefits of averaging, which increases with the statistical independence of the judges. Independence is facilitated both through crowd composition, such as selecting judges who are trained differently or have experience in different judgment environments, and through process, such as ensuring that individuals' judgments are made prior to interaction, thereby reducing the effects of anchoring (Tversky & Kahneman, 1974) and social influence (Deutsch & Gerard, 1955). In sum, with bracketing, an average will be more accurate than the judgment of the crowd's average member. Without bracketing, an average can perform no worse than the crowd's average member. The performance of averaging is thus improved when the overall level of expertise in the crowd is raised or the bracketing rate is increased.

Our discussion thus far suggests some rough conditions under which each strategy might perform well. The best-member strategy appears suited to situations in which judges differ greatly in their ability and the decision maker can capitalize on this with a valid cue to expertise. Averaging will work well when judges' estimates frequently bracket the truth so that errors cancel out. Between these two extremes, however, there will be many intermediate situations characterized by moderate levels of both dispersion and bracketing. This suggests an intermediate strategy, which we call the *select-crowd* strategy. The term *select* has a number of connotations that we mean to build upon. The crowd is select in that it is smaller than the whole crowd, it is selected or chosen, and it is of higher quality than the whole crowd (to the degree that there is dispersion in expertise).

To illustrate the idea of a select crowd, consider a general procedure that ranks all members of a group, panel, or crowd based on some cue to ability and then averages the judgments made by the most highly ranked k members. Selections of $k = 1$ and $k = N$ judges correspond to what we call the two *pure strategies*—the best-member strategy and the whole-crowd strategy, respectively. This leaves a range of strategies when k is strictly between 1 and N. Below, we investigate the entire range of k, but for now, we note that we often use $k = 5$ judges for our typical select crowd. We show below that select crowds of five members strike an ideal balance between using the best judges on the one hand and taking advantage of the error-cancelling effects of bracketing on the other.

The Role of Environment

A natural question to ask at this point is which of the three strategies for dealing with the multitude of opinions that populate the crowd is best. Clearly, there are times when seeking out the crowd's best member is the correct response. When someone goes into cardiac arrest in the middle of a restaurant, it makes sense for diners to ask, "Is there a doctor in the house?" rather than take a straw poll over what to do next. For quantity estimates, relying on the best member will yield the most accurate judgments when there are large differences in ability that can be identified by a valid cue (Davis-Stober, Budescu, Dana, & Broomell, 2014). In contrast, averaging the whole crowd works exceptionally well when small differences in ability are accompanied by frequent bracketing (Einhorn, Hogarth, & Klempner, 1977; Soll & Larrick, 2009). This illustrates that the effectiveness of a judgment strategy depends critically on the environment one faces (Csaszar & Eggers, 2013; Payne, Bettman, & Johnson, 1988). A core assumption of this article is that people often do not know their judgment environment. How much, for instance, do the economists surveyed by *The Wall Street Journal* differ in expertise? How often do their judgments bracket the truth? With a history of past performance, it is possible to start answering these questions. In the absence of history or other cues to expertise, people can only guess about the environment they are in, and because they can guess wrong, their choice of strategy may be a poor fit to the environment and yield inaccurate judgments.

Figure 1 depicts four typical judgment environments that decision makers might face. We focus on two dimensions of the environment, the dispersion in expertise and the level of bracketing, which partition the judgment space into four quadrants. Quad-

	Low dispersion in expertise	High dispersion in expertise
High bracketing	(A) Whole Crowd	(B) Select Crowd
Low bracketing	(C) Select Crowd	(D) Best Member

Figure 1. Four exemplar judgment environments and the strategies expected to perform the best in each.

rant A is a low-dispersion–high-bracketing environment. In this environment, there are small differences in expertise, the more capable judges cannot be reliably identified, and individual judgments frequently bracket the truth. Accordingly, the best-member strategy is a poor fit for this environment, and averaging the whole crowd will yield more accurate judgments. Quadrant D, in contrast, is a high-dispersion–low-bracketing environment. In this environment, there are large differences in expertise, the more capable judges can be reliably identified, and individual judgments rarely bracket the truth (which could arise, e.g., if the judges had similar backgrounds, training, or mental models and therefore tended to make the same errors; Page, 2007). This is the ideal environment for a best-member strategy, whereas averaging the whole crowd will yield poor results.

Because people often do not know their environment, it is desirable to find judgment strategies that perform well across environments—what we call *robust strategies.* These strategies free the decision maker from the difficult task of identifying the current environment and from the inaccuracy that results from guessing wrong. We propose that select crowds are robust in this sense and that the best member and the whole crowd are not. Consider first the two environments already discussed. In the low-dispersion–high-bracketing environment (see Figure 1, Quadrant A), we expect averaging the whole crowd to perform well and the best member to perform poorly. One principle of averaging is that its benefits accrue rapidly with the first set of judges but increase more slowly as additional judges are added (Hogarth, 1978). This suggests that although a select crowd may not perform quite as well as the whole crowd in Quadrant A, it could come reasonably close, and it will most likely outperform the best member. In contrast, in Quadrant D, it is important and possible to exclude the worst judges, which the select-crowd strategy accomplishes by including only a subset of the best. As a result, we expect select crowds to perform nearly as well as the best member in this environment and far better than the whole crowd. Thus, in these two environments, select crowds will perform well—nearly as well as the superior of the two pure strategies and far better than the inferior one.

Select crowds have the greatest potential in the high-dispersion–high-bracketing environment (see Figure 1, Quadrant B). Here, there are large identifiable differences in expertise, which favors choosing the best member. But judges' estimates also frequently bracket the truth, which favors averaging the whole crowd. Select crowds take advantage of both features by choosing a subset of the best judges while capitalizing on the error-cancellation benefits of averaging. We therefore expect select crowds to outperform both the best member and the whole crowd in this environment.

The final environment is a challenging one for any judgment strategy. The low-dispersion–low-bracketing environment (see Figure 1, Quadrant C) features small differences in expertise, which is unfavorable to the best-member strategy. It also features low bracketing, which is unfavorable to averaging the whole crowd. Thus, the environment is hostile to both pure strategies, neither of which we expect to perform well in an absolute sense. To the extent, however, that a select crowd takes advantage of identifiable differences in expertise, however minor, and includes at least some judgments that bracket the truth, we expect it to perform as well as if not better than both pure strategies.

In sum, we expect select crowds to perform the best in two of the four judgment environments illustrated in Figure 1 and close to the best in the remaining two. This makes select crowds a robust judgment strategy across environments, whereas the same cannot be said for the best member and whole crowd. The beauty of a robust strategy is that, because it performs well across environments, it can be applied even when the decision maker is blind to the environment he or she faces. If the decision maker has valid information about the environment, then it may pay to deviate from the select-crowd strategy, a point we return to in the General Discussion. However, even in these cases, we expect that the select crowd will not lag far behind. If an accurate guess about the environment cannot be made, then our results show that a select-crowd strategy is the most effective approach.

Simulation

To systematically examine the performance of these judgment strategies, we simulated the four environments illustrated in Figure 1. Crowds of judges were created to reflect the parameters of each environment (viz., dispersion in expertise and bracketing), and the simulated judges provided a series of numerical estimates. On each item, accuracy was defined as the AE of the estimate—the absolute value of the difference between the estimate and the truth. A judge's level of expertise was equal to the average accuracy across estimates—the mean AE (MAE). There are alternative measures of accuracy, such as correlation and mean squared error, but we prefer MAE because it is simple, has a straightforward interpretation, and is widely used in both the psychology and forecasting literatures (Armstrong, 2001). Judges with lower MAEs are considered to have more expertise because they tend to make smaller estimation errors.

Modeling the Environment

Dispersion in expertise. In all environments, the MAEs of the judges were modeled as uniformly distributed with a mean of 100.[1] We manipulated dispersion by varying the range of expertise around this average value. Our preferred metric was the coefficient of variation (CV), the ratio of the standard

[1] We examined several distributions of expertise by varying the parameters of a beta distribution (*a*, *b*). In addition to the uniform distribution (1.0, 1.0) reported in the simulations, we included an approximation to the normal distribution (2.5, 2.5), a left-skewed distribution with many bad judges (1.0, 0.5), a right-skewed distribution with many good judges (0.5, 1.0), and a bimodal distribution (0.5, 0.5). The performance of a five-person select crowd was robust across these distributions in expertise. Details are available from the authors.

deviation of a quantity to its mean. (Critically, the use of CV as the metric for dispersion allowed us in the archival data sets to compare judgments for outcomes differing in scale and crowds differing in their average levels of expertise.) For the low- and high-dispersion environments illustrated in Figure 1, CVs were set at 0.10 and 0.40, respectively. This corresponds to MAEs ranging from 83 to 117 in the low-dispersion environment and from 31 to 169 in the high-dispersion one. Stated another way, the MAE of the worst judge was 1.4 times that of the best judge in the low-dispersion environment but 5.5 times that of the best judge in the high-dispersion environment.

We supplemented Figure 1's two levels of dispersion with analyses of a larger sample of environments. We started with no dispersion in expertise (CV = 0.00) and increased CV by increments of 0.01 until the distribution became as wide as possible while holding constant the average MAE of 100. This happened at a CV of 0.55. At this level of dispersion, MAEs ranged from about 5 to 195. Thus, in the most extreme environment, the ratio of the accuracy of the worst judge to that of the best judge was about 39 to 1. (This is a very broad range that covers the large majority of CV levels we observed in the archival data sets discussed shortly.)

Bracketing. Bracketing refers to the frequency with which any two judges' estimates fall on opposite sides of the truth. It is minimized in large crowds when all judgments err on the same side of the truth (due to either shared bias or perfectly correlated errors), and it is maximized when half the judgments are on either side of the truth on every question, so that the chances that any two judges selected at random bracket the truth are about 50%. We manipulated bracketing by varying the pairwise correlation in the judges' signed errors. When two judges' errors are uncorrelated ($r = 0$), their estimates are equally likely to fall on the same side or opposite sides of the truth, so the bracketing rate is 50%. When their errors are positively correlated, the bracketing rate will be less than 50%. (Note that, although it is possible for pairs of judges to have bracketing rates over 50%, if they have opposing biases or negatively correlated errors, the *average* pairwise bracketing rate cannot exceed 50% as the number of judges in a crowd becomes very large.) For the low- and high-bracketing environments illustrated in Figure 1, we set the rates at levels near their theoretical limits, namely, at 10% (pairwise $r = .95$) and 40% (pairwise $r = .31$), respectively. We supplemented Figure 1's two levels of bracketing with an analysis of the full range of possible bracketing rates in large crowds, from 0% to 50%. Bracketing rates were held constant for all pairs of judges.

History. Within each environment, we examined the performance of the best member and select crowds after ranking the judges based on different levels of information about their expertise. For convenience, we refer to this as *history* (h). In practice, a judge's history is any cumulative record or indicator of expertise, which could include education and training, confidence, or credentials. In the simulations, history refers to a record of performance on prior judgments. We included seven levels of history. A history of zero ($h = 0$) corresponds to choosing judges based on no information whatsoever, that is, choosing judges at random. Histories of 1, 3, 5, 10, and 20 correspond to ranking judges based on their performance over the h respective prior period(s). A history of one, for example, ranks judges based only on their performance in the most recent period, a history of five ranks judges based on their average performance over the last five periods (equally weighted), and so on. With more history, it should be easier to identify the crowd's better judges. Finally, we included for comparison the performance of judges chosen with perfect knowledge of their true expertise, which is equivalent to having an infinite history for each judge.

Crossing the two levels of dispersion (CVs of 0.10 and 0.40) and two levels of bracketing (10% and 40%) created the four exemplar environments of Figure 1, each of which was examined with seven levels of history. The fuller sample of environments included the factorial combination of dispersion (CVs from 0.00 to 0.55) and bracketing rates (0% to 50%), which generated 2,856 judgment environments. To keep the computation manageable for this much larger set of environments, we held constant the history over which judges were ranked and selected at five prior periods.

Implementing the Simulation

For each run of the simulation, MAEs were randomly sampled for N judges from the appropriate uniform distribution, as described above. N indexed the size of the whole crowd, which we set at 25, 50, 100, or 200 judges. We generated 120 periods of estimates by these judges. Periods 1–20 provided the necessary history to rank and choose judges for the subsequent 100 trials (Periods 21–120), over which we evaluated strategy performance. So, for example, with history set at five periods, judges were ranked for selection in Period 21 based on their average performance over Periods 16–20. In Period 22, selection was based on their average performance over Periods 17–21, and so on. In each period, the following steps were implemented:

1. Errors were generated for each judge for the current period.[2]
2. Sample MAEs were calculated for each judge over the h preceding periods, where h is the length of the history.
3. Judges were ranked based on their sample MAEs.

[2] Our simulated judges had normally distributed signed errors with a mean of zero. (Since bracketing reflects both random-error correlation and shared bias, the assumption of mean-zero errors here is not critical to our results.) For a given judgment j, the error for judge i was set to be $M_i(\sqrt{r}Z_{sj} + \sqrt{1-r}Z_{ij})$, where Z_{sj} and Z_{ij} are standard normal deviates generated for judgment j. Z_{sj} is a shared component of error shared by all judges, whereas Z_{ij} is unique to judge i. To achieve higher levels of bracketing, less weight must be placed on the shared term and more weight on the unique term. This is accomplished by varying r, which is the average pairwise correlation in judges' signed errors and maps one-to-one to bracketing when judges are unbiased (e.g., for 10% and 40% bracketing, $r = .951$ and $r = .309$, respectively). Finally, the term M_i is a judge-specific parameter that scales up the error to match the judge's assigned MAE ($M_i = \sqrt{\pi/2}MAE_i$).

4. Strategies were implemented for $k = 1$ to N by taking the mean of the signed errors of the top k judges.[3]

After completing these steps for each period, we calculated the MAE of each strategy over 100 periods (Periods 21–120). Finally, these scores were normed against the performance of the average judge in each environment. Specifically, the MAE of a strategy was expressed as its percent improvement over the MAE of the average judge. Since the latter was fixed at 100 in each simulation, the percent improvement of the strategy was simply (100 − MAE)%. If, for example, the MAE of the strategy was 85, it was reexpressed as (100 − 85)%, or as a 15% improvement over the average judge. For the four environments of Figure 1, each simulation was run 1,000 times, beginning with a new random draw of N judges. To keep the computation manageable for the larger set of 2,856 environments, each simulation was run 200 times for crowds of 25, 50, and 100 judges and run 100 times for a crowd of 200 judges. The final results are based on averaging the percent-improvement scores over these iterations.

Results and Discussion

Exemplar environments. Figure 2 illustrates performance in the four environments pictured in Figure 1 for a crowd of 50 judges. Strategy (i.e., choice of k) is listed on the x-axes, where $k = 1$ is the best member, $k = N$ is the whole crowd, and $1 < k < N$ are select crowds. The performance of each strategy is on the y-axes, with higher values indicating better performance. The panels are set up as in Figure 1, with rows representing the bracketing rate and columns the dispersion in expertise. Each panel shows performance for seven levels of history.

We note three important patterns in Figure 2: the effect of environment on select-crowd performance, the relatively similar performance of select crowds of different size in the vicinity of $k = 5$ judges, and the role of history. First, the results conform to the pattern anticipated by our earlier analysis of the relationship between performance and dimensions of the environment. Select crowds excel when both dispersion and bracketing are high (see Figure 2B) or when both are low (see Figure 2C), the best member excels when dispersion is high and bracketing is low (see Figure 2D), and the whole crowd does well when dispersion is low and bracketing is high (see Figure 2A). Although the select crowd does not perform as well as a pure strategy in Figures 2A and 2D, its performance is very close.

Second, Figure 2 reveals that the performance of select crowds in the range of three to eight judges is relatively similar, depending on environment and history. With a history of five periods, for instance, the performance of select crowds in this range was 29.7%–39.4% in Figure 2A, 65.0%–66.5% in Figure 2B, 10.6%–11.3% in Figure 2C, and 56.3%–62.4% in Figure 2D. The performance of a five-person select crowd in these environments was 35.8% (see Figure 2A), 66.6% (see Figure 2B), 11.1% (see Figure 2C), and 60.1% (see Figure 2D). In other words, although choosing slightly more or fewer judges may maximize performance of the select crowd in specific environments, a five-person select crowd by and large delivers most of the benefit. Accordingly, when one is unsure of the environment, a default choice of five judges is a reasonable rule of thumb.

Third, Figure 2 illustrates that the performance of select crowds increases with more history. This is revealed in Figures 2B and 2D by the large difference between the lowest and highest curves, reflecting the gap between picking a select crowd at random and one according to true skill, respectively. However, as more history is collected, improvements in performance grow smaller, as indicated by the closer spacing of the curves based on longer histories. In other words, there are diminishing returns to collecting more information to rank and choose judges. An interesting implication of this is that a short history is sufficient to discriminate ability when dispersion is high. In the high-dispersion environment (CV = 0.40), the correlations between the ranking of the judges by their true expertise and their rankings estimated with one, five, and 10 periods of history were .42, .77, and .86, respectively. A history of five periods was an excellent cue to expertise in this environment, and even one period of history provided useful information.

Moreover, short histories are sufficient when dispersion is low, but for a different reason. In this case, history is a less valid cue to expertise. In the low-dispersion environment (CV = 0.10), the correlations between the ranking of the judges by their true expertise and their rankings estimated with one, five, and 10 periods of history dropped to .11, .27, and .37, respectively. The reason for this is straightforward: When there is little variation in ability among the judges, there is little that history can explain. As a result, the benefits of collecting longer histories in low-dispersion environments are minor because the best and worst judges perform at similar levels. This is revealed in Figures 2A and 2C by the small difference in performance between the lowest and highest curves. Because realized performance is always bounded between that of randomly ranking the judges (lowest curve) and perfectly ranking them (highest curve), which is narrow when there is low dispersion, the incremental improvement in ranking that comes with more history does not have much benefit.

In sum, short histories are effective in high-dispersion environments because they provide valid cues to expertise, and they are sufficient in low-dispersion environments because the total return that comes with longer histories is limited. Thus, when the environment is unknown, the expected gain from collecting longer histories is quite small. This is fortunate for decision makers because longer histories of past performance are often either unavailable or prohibitively expensive. Moreover, if expertise changes over time, any ranking based on a long history would be obsolete and ineffective.

A wider range of environments. We turn next to our analysis of the set of 2,856 environments in which dispersion in expertise (CV) ranged from 0.00 to 0.55 and bracketing rates ranged from

[3] Because the median is often used as a measure of central tendency in place of the mean, particularly for highly skewed distributions, we also evaluated its performance when used as the judgment of a select crowd and the whole crowd. For brevity, we omit the details on the performance of medians, but we note the following. On average, the median judgment performed better than the mean judgment, but this difference was apparent only for high values of k. For instance, in the high-dispersion–high-bracketing environment of Figure 1, the difference in performance between the median and mean judgment of 50 judges was 8.8 percentage points, whereas the difference for five judges was 1.1 percentage points. This pattern reveals another advantage of the select-crowd strategy: Its performance is less sensitive to the choice of median or mean than is the performance of the whole crowd.

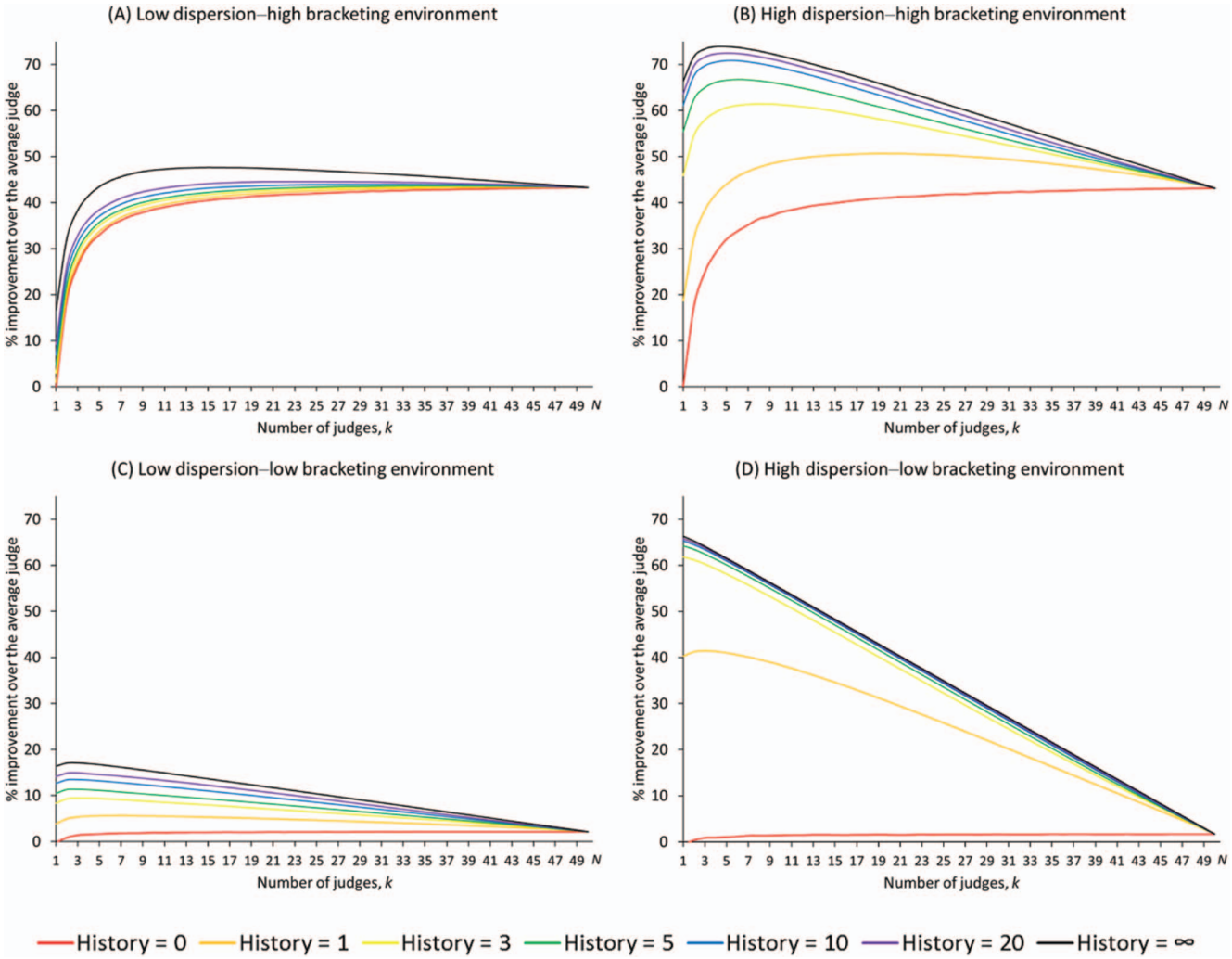

Figure 2. Performance of judgment strategies for a simulated crowd of 50 judges. The performance of the best member is indicated at $k = 1$, of the whole crowd at $k = N$, and of select crowds at $1 < k < N$. Curves are shown for judges ranked and selected based on performance over seven levels of history. The lowest curve in each graph (History = 0) corresponds to choosing k judges at random, and the highest curve (History = ∞) corresponds to choosing k judges according to their true skill based on a full history.

0% to 50%. Figure 3 provides contour maps of the performance of the best member, the whole crowd, and a five-person select crowd (where $N = 50$ judges and $h = 5$ periods). In each panel, performance is plotted against dispersion on the *x*-axis and bracketing on the *y*-axis. Regions with deeper shading indicate higher levels of performance. Although the figure indicates that all three strategies perform better than choosing judges at random, it illustrates why we claim select crowds are a robust judgment strategy. In the left panel of the figure, the performance of the best member varied mostly with changes in dispersion (indicated by the vertical bands of shading that darken as dispersion increases). The best-member strategy is appropriate when dispersion in expertise is high, and its performance is largely unaffected by bracketing rates. In the right panel, the performance of the whole crowd varied mostly with changes in bracketing (indicated by the horizontal bands of shading that darken as bracketing increases). The whole-crowd strategy is appropriate when bracketing rates are high, and its performance is largely unaffected by dispersion in expertise. Both pure strategies had large regions in which they performed poorly. In contrast to these strategies, select crowds take advantage of changes in *both* dimensions of the environment. In the middle panel, the performance of a five-person select crowd varied with changes in both dispersion in expertise and bracketing rates (as indicated by the concentric bands of shading that darken as both variables increase). Figure 3 clearly illustrates that the select-crowd strategy performs well across a greater range of judgment environments than either the best member or the whole crowd.

Using the same data, Figure 4 identifies the regions in which the best member, the whole crowd, and the five-person select crowd perform best (cf. Figure 2, Soll & Larrick, 2009, which defines regions for the best member and averaging for the case of two judges). The football-shaped region in which the select crowd

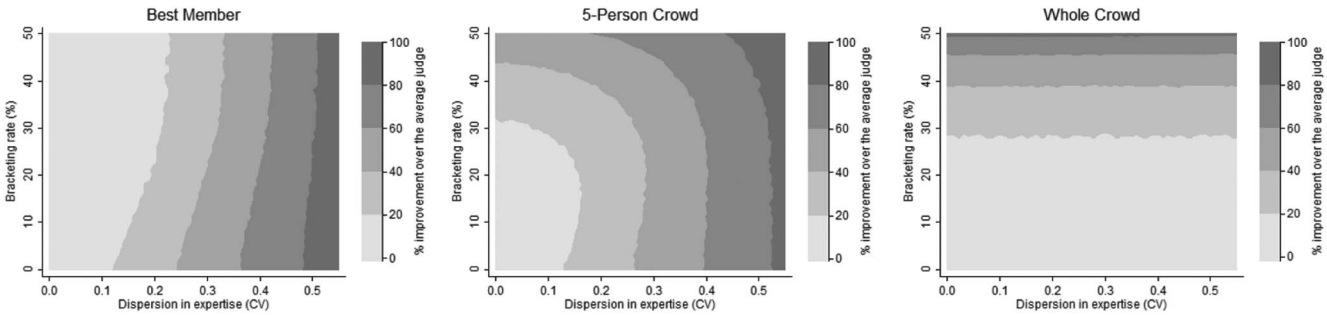

Figure 3. Contour maps of performance across 2,856 simulated judgment environments for three judgment strategies. Five trials of history were used to rank and select judges ($N = 50$). Darker shades of gray indicate greater percent improvement over the average judge. CV = coefficient of variation.

performs best is quite large, and it includes many environments people are likely to encounter. In the archival data we discuss next, we found that CVs in the 0.20–0.40 range and bracketing rates between 20% and 40% were very common; select crowds do well in this space. Figure 4 also illustrates why select crowds are robust. Because its region of dominance lies between those of the two pure strategies, the select crowd is nearly always either the best strategy itself or the second best strategy. Unlike the pure strategies, select crowds rarely perform the worst.

Figure 5 presents another view of the results. Here, we have plotted the performance of the strategies for crowds of 25, 50, 100, and 200 judges. On the *x*-axes, we show results for values of *k* up to 25 judges and for values close to *N*. Performance is presented on the *y*-axes, but in a modified metric (*relative performance*), calculated as follows. In each of the judgment environments, we determined which strategy delivered the greatest improvement over the average judge (i.e., the optimal value of *k*).[4] We then measured the performance of all other strategies relative to the best strategy. For example, $k = 1$ was the top-performing strategy in the environment corresponding to dispersion of 0.40 and bracketing of 10%, with a 64.5% improvement in accuracy over the average judge ($N = 50$ judges, $h = 5$ periods). The improvement of $k = 5$ in this environment was 59.9%. So, the relative performance of $k = 5$ compared to the best-performing strategy in this environment was .93 (= 59.9/64.5). This calculation was repeated for all levels of *k* in each environment and then averaged over the set of all judgment environments for each crowd size. Figure 5 plots the mean level of relative performance for each level of *k*, as well as the variability in relative performance across environments, as indicated by lines for the 5th and 95th percentiles.

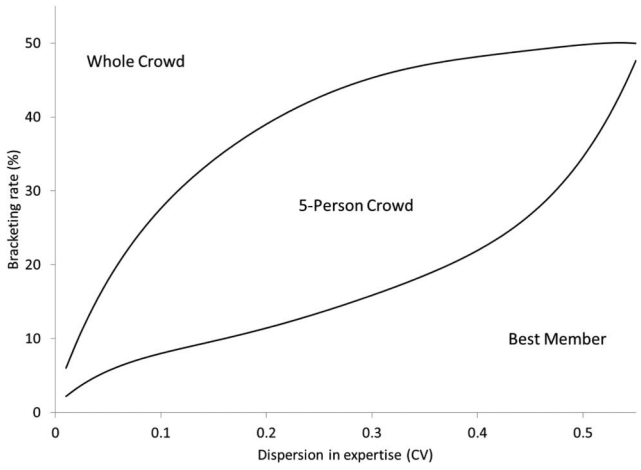

Figure 4. Best-performing strategy for each simulated judgment environment with $N = 50$ judges ranked and selected based on five periods of history. With less (more) history available to select judges, the curves rotate clockwise (counterclockwise). CV = coefficient of variation.

We note several insights from Figure 5. Select crowds performed well regardless of the overall size of the crowd from which they were drawn. The mean relative performance of a five-person select crowd ranged from .90 for $N = 25$ to .94 for $N = 200$. As *N* increases, more judges are needed to maximize the relative performance of select crowds (e.g., maximum relative performance was delivered by a 12-person select crowd for $N = 200$), but five judges performed nearly as well as the optimal crowd size. Next, select crowds not only delivered high mean relative performance, they also had less downside risk compared with the best member and the whole crowd. This is indicated by the dotted lines for the 5th percentile of relative performance in the figure. In Figure 5B ($N = 50$), for instance, the relative performance of a five-person select crowd exceeded .77 in 95% of the environments. The corresponding levels for the best member and the whole crowd were .02 and less than .01, respectively. In other words, when select crowds were not the best strategy, they did not trail the best by much, unlike the best member and the whole crowd. Finally, for select crowds, there is little tradeoff between mean relative performance and variability in relative performance. As indicated by the figure, the 5th percentiles of relative performance were higher (i.e., better) for slightly larger select crowds. In Figure 5B ($N = 50$), mean relative performance was maximized by a six-person select crowd ($M = .933$, 5th percentile = .812), whereas the 5th percentile of relative performance was maximized by an eight-person select crowd ($M = .927$, 5th percentile = .853). This pattern was consistent across the panels of Figure 5. Given the consistently solid performance of select crowds in the range of four to eight judges, we reiterate selecting five judges as an effective rule of thumb.

[4] We omitted environments in which the improvement of the best strategy over the average judge was less than 1%. These comprised less than 0.5% of the total, leaving us with 2,844 judgment environments.

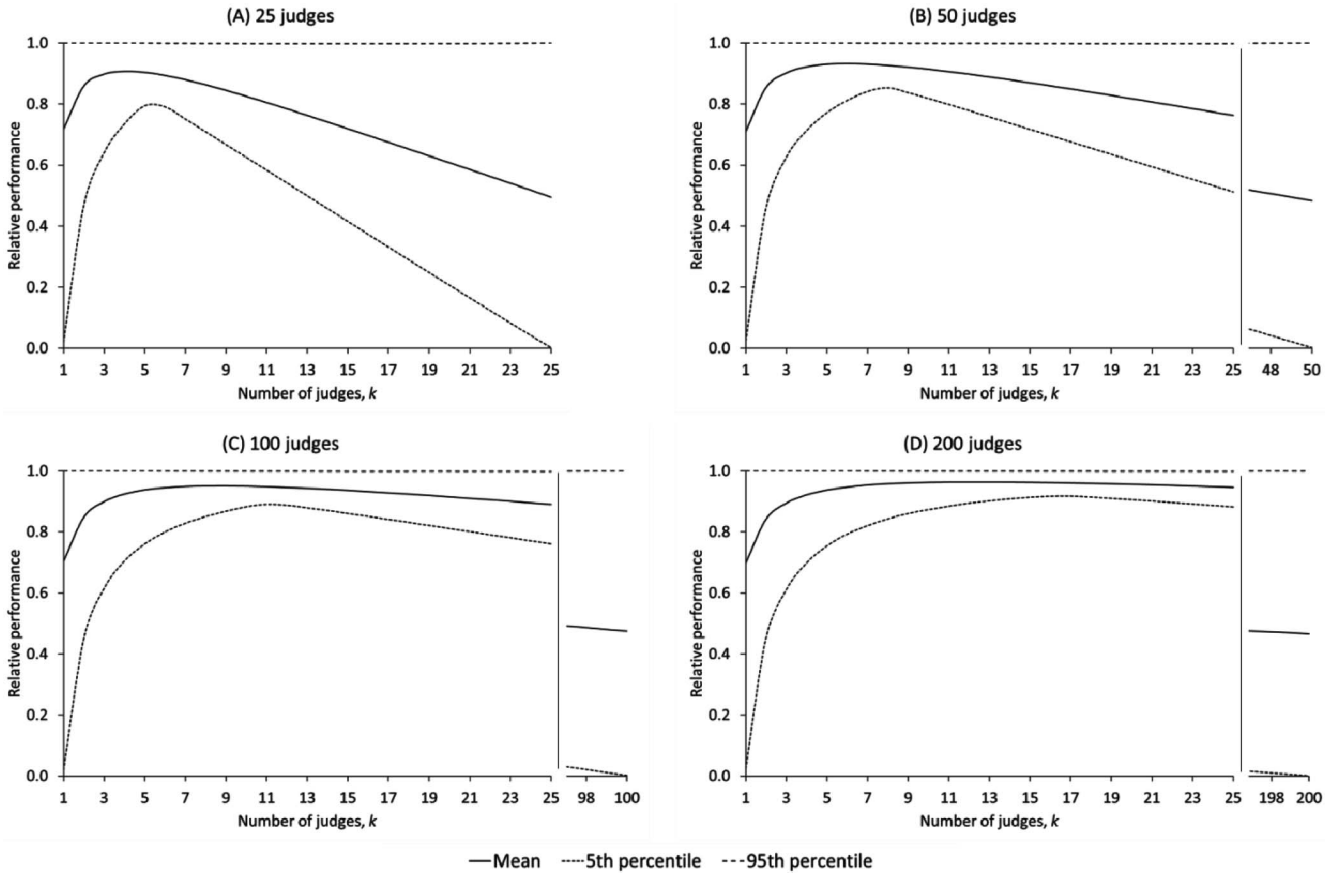

Figure 5. Distribution of the mean, 5th percentile, and 95th percentile of relative performance based on five periods of history across 2,844 simulated judgment environments for crowds of 25, 50, 100, and 200 judges. The *y*-axes plot the percent improvement of the strategy relative to that of the best-performing strategy in each environment, so higher values are better. The line for the 5th percentile indicates the level of relative performance that each strategy exceeded in 95% of the environments. Relative performance is shown for values of k up to 25 judges and values of k close to N. The best member is indicated at $k = 1$, the whole crowd at $k = N$, and select crowds at $1 < k < N$.

Validation

Having established through simulation that select crowds are a viable alternative to the best member and the whole crowd for improving judgment, we turn next to validating these findings with archival data. Based on the simulations, we expected the average performance of a five-person select crowd to significantly exceed that of the best member and the whole crowd. We also expected the select crowd to rank first or second in performance among the three strategies in the large majority of cases.

Data Sets

We assembled two types of data, one consisting of estimates made by participants in laboratory experiments (*experimental data*) and the other consisting of forecasts made by professional economists (*economic data*). The experimental data comprised numerical estimates made by participants in published research conducted for other purposes (Larrick, Burson, & Soll, 2007; Mannes, 2009; Moore & Klein, 2008; Moore & Small, 2008; Soll & Larrick, 2009; Soll & Mannes, 2011). These articles featured multiple studies, multiple samples within studies, and multiple domains (e.g., temperatures, distances, etc.) for which participants provided private and independent responses to questions with known answers. The number of judges in these studies ranged from 15 to 413, and the number of questions ranged from 10 to 38. Altogether, these articles provided 40 unique sets of nearly 33,000 numerical estimates made by 1,400 people.

The economic data comprised forecasts of economic indicators by professional economists. These are publicly available from the Federal Reserve Bank's *Survey of Professional Forecasters*.[5] We used forecasts for seven economic indicators: the consumer price index, the rate on the 3-month Treasury Bill, the rate on the 10-year Treasury Note, the yield on Moody's AAA corporate bond, nominal gross domestic product, housing starts, and

[5] http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/

the unemployment rate. (Results for the last three indicators were calculated for two separate eras reflecting a change in the survey's administrator and participants that occurred in 1992.) Forecasts for six horizons are elicited from economists once per quarter: the quarter just ended (Horizon 1), the current quarter (Horizon 2), and the four prospective quarters (Horizons 3–6). For example, in the survey administered in the first quarter of 2013 (2013:Q1), economists made forecasts for 2012:Q4 (Horizon 1), 2013:Q1 (Horizon 2), 2013:Q2 (Horizon 3), 2013:Q3 (Horizon 4), 2013:Q4 (Horizon 5), and 2014:Q1 (Horizon 6). We included the forecasts made for five horizons (2–6) in our analyses. The median number of judges per survey was 35. Altogether, the economic data comprised nearly 160,000 forecasts from 1968 to 2012 for 50 sets of data (where a set is defined by indicator–era–horizon). We extracted the realized values of the economic indicators (first vintage) from the *Real-Time Data Set for Macroeconomists* (see footnote 5).

For each of our 40 sets of experimental data and 50 sets of economic data, we compared the performances of the best member ($k = 1$), the whole crowd ($k = N$), and select crowds ($1 < k < N$). For the best member and select crowds, judges were ranked and selected based on six levels of historical performance. Details on these variables and the steps used to calculate performance are presented in the Appendix.

Results and Discussion

Figure 6 plots the performance of the best member, select crowds, and whole crowd in the experimental and economic data, averaged over environments. As in the simulations, strategies are listed on the *x*-axes, with the best member and the whole crowd indicated at $k = 1$ and $k = N$. On the *y*-axes is the percent improvement of the strategy over the average judge, so higher values indicate superior performance.

In both data sets, select crowds were clearly superior to the best member and, with sufficient history, proved superior to the whole crowd as well. In the experimental data, the average improvement of a five-person select crowd chosen based on five periods of

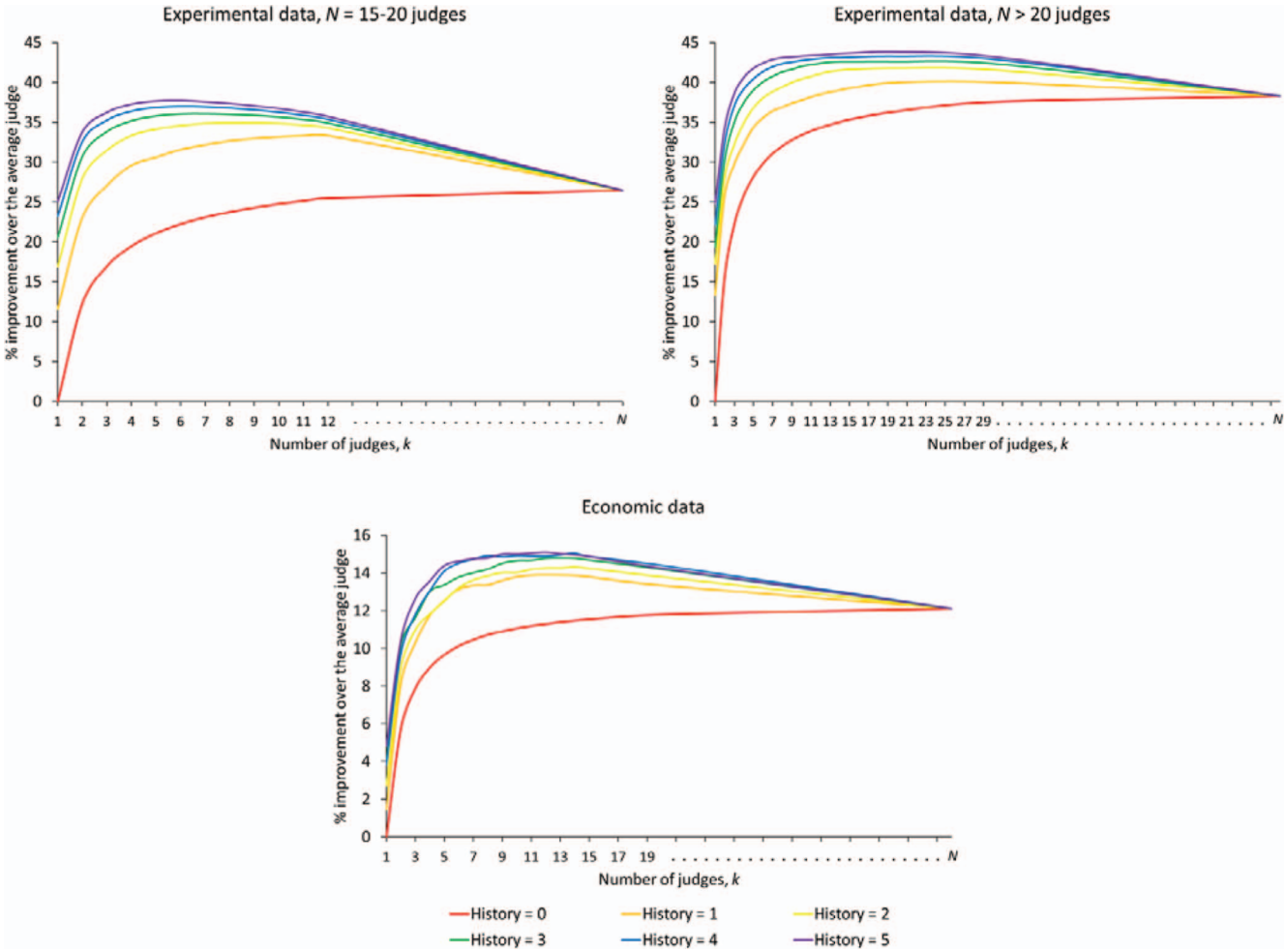

Figure 6. Performance of the best member ($k = 1$), select crowds ($1 < k < N$), and whole crowd ($k = N$) in the experimental and economic data. The experimental data is separated into the 20 sets with $N = 15$–20 judges (top left) and the 20 sets with $N > 20$ judges (top right). Selections based on six levels of history are provided. Performance for omitted values of k (in ellipses) is interpolated.

history was 39.7% (SD = 15.1%), which was significantly better than the average improvement of both the best member (M = 25.0%, SD = 19.9%), $t(39) = 6.13$, $p < .001$, Cohen's $d = 0.84$, and the whole crowd (M = 32.4%, SD = 14.2%), $t(39) = 2.92$, $p = .006$, $d = 0.51$. Similarly, in the economic data, the performance of the five-person select crowd (M = 14.4%, SD = 7.6%) was significantly better than that of both the best member (M = 4.8%, SD = 6.7%), $t(49) = 10.58$, $p < .001$, $d = 1.35$, and the whole crowd (M = 12.1%, SD = 5.7%), $t(49) = 3.52$, $p < .001$, $d = 0.34$.

Table 1 testifies to the robustness of the select-crowd strategy. Here, for each of the 40 sets of experimental estimates and 50 sets of economic forecasts, we ranked the performance of the best member, the whole crowd, and a five-person select crowd. There were significant differences in the rankings of the judgment strategies for both the experimental and economic data (ps < .001 by Fisher's exact tests). In the experimental data, the select crowd ranked first in performance in 21 of the 40 sets of estimates (53%), compared with 14 sets (35%) for the whole crowd and five sets (13%) for the best member. Equally impressive, the select-crowd strategy ranked second in 18 of the 19 sets in which it was not ranked first. In the economic data, the select crowd ranked first in performance in 34 of the 50 sets of forecasts (68%), compared with 15 sets (30%) for the whole crowd and one set (2%) for the best member. The select-crowd strategy ranked second in 14 of the 16 sets in which it was not ranked first. These results illustrate the robustness of the select-crowd strategy and support what the simulations predicted in Figures 4 and 5—a select crowd is often the best and rarely the worst.

We conclude the following from our analyses of the simulated judgment environments and the archival data in the first part of this article: (a) Averaging the opinions of a select crowd of five judges is a very robust judgment strategy. It takes advantage of expertise when experts can be identified and of bracketing when judges are independent. As a result, it is suited to a wide range of judgment environments, whereas the best member and the whole crowd perform well only in narrower ranges. (b) In those environments in which the select crowd does not perform the best, it is usually a close second, and it rarely performs the worst. (c) Decision makers do not need a lot of history for the select-crowd strategy to work well. When dispersion in expertise is high, one to five observations of past performance are sufficiently diagnostic for selecting judges; when dispersion is low, the total returns to history are marginal in the first place, so short histories are sufficient. (d) The strategy performs well regardless of the size of the crowd from which the judges are drawn and whether the select crowd's mean or median judgment is used for estimation (see footnote 3).

Table 1
Counts for Ranked Performance of the Best Member, Whole Crowd, and Select Crowd in the Experimental (N = 40) and Economic (N = 50) Data Sets

Strategy	1st	2nd	3rd
		Rank in experimental data	
Best member	5	9	26
Whole crowd	14	13	13
5-person select crowd	21	18	1
		Rank in economic data	
Best member	1	9	40
Whole crowd	15	27	8
5-person select crowd	34	14	2

Note. The best member and select crowd were ranked and selected based on five periods of history.

Having established through simulations and analysis of archival data the promise of select crowds for improving judgment, we turn in the second part of the article to three experiments in which we examined people's perceptions and actual use of the strategy.

The Psychology of Judgment Aggregation

Research has shown that people are skeptical of purely statistical approaches to decision making, including the use of linear models like averaging (Kleinmuntz, 1990). People are reluctant to take a simple average of opinions, for example, because it treats the opinions of experts and nonexperts alike, which in their eyes is a recipe for mediocrity (Larrick & Soll, 2006). In a preliminary experiment, we asked 46 women and 50 men (M_{age} = 35.5 years) from an online panel to evaluate methods of forecasting attendance at a hypothetical film series. The participants were introduced to five members of a committee for this film series—Amy, Brad, Carrie, Doug, and Eric—who varied in their ability to accurately forecast attendance at upcoming events. On average, Amy (the most accurate forecaster) erred by only 10 people per film—sometimes high, sometimes low—Brad by 20 people, Carrie by 30 people, Doug by 40 people, and Eric by 50 people. When asked for their best estimate of the error they expected from averaging the forecasts of all five committee members, 70% of the participants (n = 67) responded with 30 people. In other words, a majority of the participants believed averaging would perform no better than the average committee member. As we have discussed, this is a mistake—averaging performs no better than the average judge only if there is no bracketing. With bracketing, averaging will always outperform the average judge.

There is also an undeniable draw to searching for the best member of a group. This may result from people's natural tendency to focus on dispositional explanations for performance instead of situational ones (Ross & Nisbett, 2011). So while *The Wall Street Journal* celebrates individual excellence in forecasting, it says little about how well the average forecast of its panel of economists performs survey after survey. Moreover, people are confident in their ability to identify the experts, which discourages them from averaging. Unfortunately, they are usually overconfident. Although some studies have found positive associations between perceived and actual expertise (e.g., Bonner, Baumann, & Dalal, 2002; Bottger, 1984; Littlepage, Robison, & Reddington, 1997, Study 3), these correlations are not large, and other studies have found null or negative relationships (e.g., Henry, 1995; Littlepage et al., 1997, Studies 1 and 2; Littlepage, Schmidt, Whisler, & Frost, 1995; Miner, 1984). A fair assessment may be that people are reliably capable of identifying the better members of a crowd but not its best member (Bonner, 2004; Miner, 1984). In practice, people are influenced by imperfect cues to expertise such as talkativeness (Littlepage et al., 1995), confidence (Anderson, Brion, Moore, & Kennedy, 2012; Sah, Moore, & MacCoun, 2013), or how much information a judge possesses (Budescu, Rantilla, Yu, & Karelitz, 2003).

Taken together, these factors lead people away from averaging toward a best-member strategy. In experiments on advice taking, for instance, people state their own opinions and subsequently have the chance to revise them aided by one or more opinions (advice) provided by others (Bonaccio & Dalal, 2006; Yaniv, 2004). Instead of simply averaging the two opinions, people typically overweight their own (Harvey & Harries, 2004; Mannes, 2009; Soll & Mannes, 2011; Yaniv & Kleinberger, 2000). In one representative study, people chose to stay with their original belief 36.1% of the time, deferred entirely to the advisor 10.0% of the time, and took a simple average of the two opinions (defined as weights between 40% and 60%) only 17.5% of the time (Soll & Larrick, 2009). People preferred one opinion or the other, depending on which they believed was closer to the truth, to a simple average.

Other research reveals a similar pattern. In experiments in which individuals must arrive at a judgment based on the opinions of others, people gravitate toward a best-member strategy (Birnbaum & Mellers, 1983; Budescu et al., 2003; Fischer & Harvey, 1999). In one study, Soll and Mannes (2011) rewarded participants for making accurate judgments based on the input of two judges. The participants were willing to use a simple average of the two opinions (defined as weights between 40% and 60%) when they viewed the judges as equally knowledgeable. When one judge was more accurate than the other, averaging declined in favor of relying on the more knowledgeable judge. The study identified perceived expertise as a fundamental driver of how people weight the opinions of others.

This evidence suggests that when given a choice between the two pure strategies—the best member and the whole crowd—people prefer the former. What if a select crowd is added to the choice set? We believe the select-crowd strategy will have intuitive appeal for many people. It allows those inclined to seek out expertise to do so while hedging their inability to identify the best member perfectly; they need only identify and exclude the poorer judges. People who (mistakenly) believe that averaging performs no better than the crowd's average judge will nonetheless appreciate that the average judge of a select crowd is smarter than the average judge of the whole crowd. We observed this in the aforementioned film-series study. When asked to estimate the expected error of averaging the forecasts of just Amy and Brad—the two best forecasters on the committee, who erred on average by 10 and 20 people, respectively—59 of the 96 participants (61%) responded with 15 people, the average performance of Amy and Brad. So even if people do not appreciate the value of bracketing, select crowds will have appeal because people view them as collectively smarter than the whole crowd.

The following studies examined people's intuitions about these judgment strategies. Experiment 1 examined perceptions of the effectiveness of select crowds compared to the pure strategies of the best member and the whole crowd. Experiment 2 tested the use of different strategies in a forecasting task for real stakes. Experiment 3 explored people's reasons for their attraction to select crowds.

Experiment 1

In this study, we presented people with a simple forecasting scenario and asked them to evaluate the accuracy of different judgment strategies. We expected them to view the select-crowd strategy more favorably than both the best-member and whole-crowd strategies.

Method

Six hundred and twenty-nine adults from a national panel in the United States (M_{age} = 45.0 years, 387 female) completed a short online questionnaire in exchange for a small payment. All passed a one-question attention check to verify that they had read instructions before being assigned to their condition and viewing the stimuli. Participants were shown the following information:

> Every year *The Wall Street Journal* surveys a panel of 50 economists. The journal asks these economists to make forecasts of several economic statistics, including unemployment, the inflation rate, etc. After one year the journal summarizes the performance of each economist based on an overall measure of how close they were to the actual values of those economic statistics. Imagine that you could earn money making accurate forecasts of the 10-year Treasury Note rate. This is the rate of interest the U.S. government pays for 10-year loans (the annualized rate as of June 2013 was 2.50%). Because you are not a professional economist, you choose to rely on *The Wall Street Journal*'s panel of economists to inform your own forecast of the Treasury Note rate one year from today.

About half of the participants (n = 312) were asked to rate the accuracy of the following five judgment strategies (1 = *not at all accurate* to 7 = *extremely accurate*): (a) the forecast of a randomly chosen economist, (b) the average forecast (the mean) of the entire panel of economists, (c) the forecast of the most accurate economist from last year, (d) the forecast of the most accurate economist over the past 5 years, and (e) the average forecast (the mean) of the five most accurate economists from last year. The remaining 317 participants were asked to rank the same strategies (without ties) based on their expected accuracy (1 = *most accurate*, 5 = *least accurate*). For both the ratings and rankings, the strategies were presented simultaneously on the computer screen in one of five orders based on a Latin-square design. Participants were randomly assigned to their task and an order.

Results

Table 2 presents the mean rating for each strategy and the difference in means for all comparisons. Our analysis of the Latin square treated order as a between-subjects factor and strategy and the listed position of each strategy as repeated measures. This design allows one to test the main effects of each factor and the interaction between the within-subject factors (see Winer, Brown, & Michels, 1991, p. 703). The effects of order, $F(4, 307) = 0.49$, $p = .74$, and position, $F(4, 1228) = 1.19$, $p = .31$, on rated accuracy were not significant, and there was no interaction between strategy and position, $F(12, 1228) = 0.71$, $p = .75$. As expected, there was a main effect of strategy, $F(4, 1228) = 161.52$, $p < .001$, $\eta_p^2 = .34$. Ratings for the select-crowd strategy (averaging the top five economists from last year) and one best-member strategy (the most accurate economist over the last 5 years) received the highest evaluations, which were statistically equivalent ($t(311) = 1.01$, $p = .31$). The select-crowd strategy was rated significantly higher than the other best-member strategy (the most accurate economist from last year; $t(311) = 6.78$, $p < .001$) and

Table 2
Ratings of Judgment Strategies in Experiment 1

Strategy	*M*	*SD*	Difference in means 1	2	3	4	5
1. Random economist	3.24	1.37	—				
2. Average of all economists	4.71	1.22	1.46***	—			
3. Most accurate economist last year	4.60	1.28	1.35***	−0.11	—		
4. Most accurate economist last 5 years	5.04	1.22	1.79***	0.33***	0.44***	—	
5. Average of 5 most accurate economists last year	5.11	1.20	1.86***	0.40***	0.51***	0.07	—

Note. $N = 312$.
*** $p < .005$ (Bonferroni-adjusted, $\alpha_{FW} = .05$).

the whole-crowd strategy (averaging the entire panel of economists; $t(311) = 5.79$, $p < .001$).

Table 3 summarizes the ranking of judgment strategies. An analysis of variance for ranked data (Winer et al., 1991) revealed a significant effect of strategy, $\chi^2(4) = 366.78$, $p < .001$. The select crowd had the lowest mean rank (i.e., highest perceived accuracy) and was ranked first by 122 participants (38.5%), followed by the most accurate economist over the last 5 years (90 participants, 28.4%). There was no effect of listed position on the rankings, $\chi^2(4) = 6.23$, $p = .18$, nor was there an effect of order (because mean ranks could not differ between participants).

Discussion

In both their ratings and rankings, the participants in this experiment expressed favorable impressions of a judgment strategy that averaged the forecasts of a select group of high-performing economists. The results reveal an intuitive appeal to select crowds. However, we deliberately made the availability of this strategy salient in this study, so it remains to be seen if unprompted decision makers also take advantage of select crowds in a task with real stakes. This was addressed in Experiment 2, in which participants made a series of predictions based on actual forecasts from a panel of economists and were rewarded for accuracy.

Experiment 2

In this section, we describe the results of an experiment in which participants made a series of forecasts with advice from 11 economists featured in *The Wall Street Journal*'s semiannual surveys. Participants were rewarded for the accuracy of their forecasts, and our primary dependent measure was whether they chose to average all 11 economists' forecasts, average the forecasts of a select crowd, or rely on a single economist. We expected participants to prefer a best-member strategy over the whole-crowd strategy in the absence of other options, but to choose select crowds significantly more often than the pure strategies when the option was available. Moreover, we expected those using the select-crowd strategy to make lower forecast errors than those using the pure strategies.

Method

Sample. Eighty students from a university in the southeastern United States ($M_{age} = 20.6$ years, 42 female) participated in exchange for $5.00 and a bonus based on the performance of the economists they chose in the study ($M = \$3.53$, $SD = \$0.22$).

Design and materials. Participants were assigned to one of four conditions created by crossing two between-subject factors, choice and feedback. A total of 24 trials in each condition represented a within-subject factor. On each of these trials, participants in the *low-choice* condition could choose one economist (best-member strategy) or average the forecasts of all 11 economists (whole-crowd strategy). Participants in the *high-choice* condition could choose any number of economists—one economist (best-member strategy), all 11 economists (whole-crowd strategy), or any subset of two to 10 economists (select-crowd strategy). Moreover, participants in both conditions were allowed to make different choices on different trials. Participants in the *low-feedback* condition received feedback about performance on the prior trial only. This feedback consisted of the 11 economists' forecasts of the criterion, the realized value of the criterion, and the economists' AEs. The feedback also included the participant's AE on the prior trial, which was determined by the accuracy of the (average) forecast of the participant's chosen economist(s). Participants in the *high-feedback* condition received the same information as those in the low-feedback condition, but they also received infor-

Table 3
Ranking of Judgment Strategies in Experiment 1

Strategy	*M*	*SD*	Frequency ranked 1st	2nd	3rd	4th	5th
1. Random economist	4.35	1.23	24	14	18	31	230
2. Average of all economists	2.86	1.23	58	68	73	95	23
3. Most accurate economist last year	3.17	1.09	23	68	86	111	29
4. Most accurate economist last 5 years	2.45	1.20	90	74	88	49	16
5. Average of 5 most accurate economists last year	2.15	1.21	122	93	52	31	19

Note. $N = 317$.

mation about cumulative performance. This consisted of each economist's MAE over all prior trials and the participant's MAE.

Eleven economists were chosen from *The Wall Street Journal*'s semiannual survey of forecasters who were present for all 24 surveys conducted from January 1994 to July 2005. Each provided forecasts of the interest rate on 3-month U.S. Treasury Bills (T-Bills) 6 months from the survey date. For reference, the average economist on this panel had an MAE of 0.55 ($SD = 0.10$) during these surveys (i.e., 55 basis points), the panel's dispersion in expertise (CV) was 0.19, and the average bracketing rate was 27%. The economists and forecasts were identical for each participant, differing only in the order in which they were presented on the screen. (Within each condition, the 20 participants were assigned a unique presentation order of the economists, which was constant throughout the experiment. Because order was not of substantive interest, we do not discuss it further.)

Procedure. Each trial began with the presentation of a table with the 11 economists' names in the first column and their forecasts of the T-Bill rate for the most recent period, the realized T-Bill rate, and their AEs in subsequent columns (including updated MAEs for participants in the high-feedback condition). Below the economists' names was a row for the "Average Forecast" of the 11 economists. The last row of the table was reserved for feedback about the participant's performance. After reviewing this information, participants chose (by checking a box) which economist(s) they wanted to rely upon for the next period's forecast. Participants in the low-choice condition selected either one economist or the "Average Forecast" of the 11 economists. Participants in the high-choice condition could do the same or select any subset of economists whose forecasts they wished to average. (Note that participants were not making their own forecasts; they were relying on the forecast or average forecast of their chosen economists. These forecasts, moreover, were unknown to the participants at the time of selection.) The trial ended once participants chose their economists, and the subsequent trial began with an updated table. Once finished with the 24 trials, participants answered questions about the expected performance of different forecasting strategies and their beliefs about bracketing. They then learned their bonuses, were paid, and were dismissed.

Results

Strategy selection. We coded selecting one economist as a best-member strategy, two to 10 economists as a select-crowd strategy, and all 11 economists as the whole-crowd strategy. Figure 7 plots the number of economists selected on the *x*-axis and the frequency of that choice on the *y*-axis for Trials 2–24.[6] In the low-choice condition, the typical preference for the best member over the whole crowd is starkly apparent: People chose one economist 82.5% of the time and the whole crowd 17.5% of the time. The median participant in this condition chose the whole crowd on only two trials, nine participants never chose the whole crowd, and only four participants chose the whole crowd more than half the time. Participants in the high-choice condition chose one economist 13.6% of the time, the whole crowd 25.0% of the time, and select crowds 61.4% of the time (ranging from two to seven economists). Not surprisingly, there was considerable experimentation in this condition, with 19 of the 40 participants trying all three strategies at some point. The top three strategies were a select crowd of two economists (29.6%), the whole crowd (25.0%), and a single economist (13.6%).

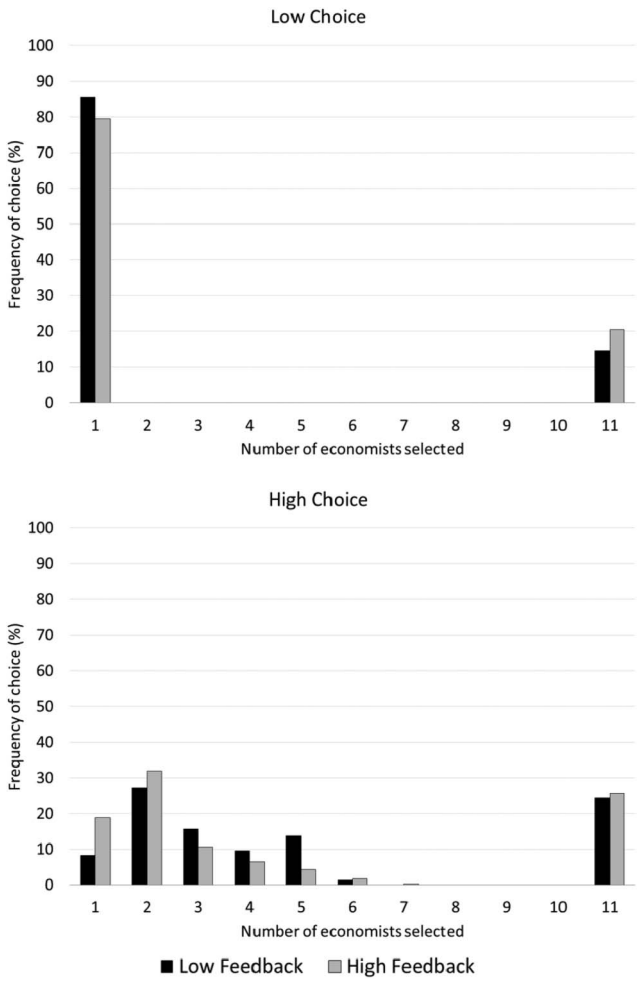

Figure 7. Frequency of judgment strategies across trials in Experiment 2 by condition.

Comparing the top and bottom panels of Figure 7 indicates that when given the option, people shifted their choices from a best member to select crowds. The effects of feedback were less obvious, so we turned to multinomial regression. Analyses were conducted separately for the low-choice condition, in which the choice of the best member or the whole crowd was modeled as a categorical dichotomous dependent variable, and the high-choice condition, in which the choice of the best member, a select crowd, or the whole crowd was modeled as a categorical trichotomous dependent variable. Feedback and a linear effect of trial were included as predictors, and standard errors were adjusted to account for the clustering of observations within

[6] On the first trial, there was no information about the economists' accuracy. Thus, choosing a best member or a select crowd was essentially choosing judges at random. As a result, we excluded the first trial from our analysis.

person.[7] Feedback had no significant effect on the choice between the best member and the whole crowd in the low-choice condition, $\chi^2(1) = 0.62$, $p = .43$. In the high-choice condition, participants receiving high feedback (i.e., information about both recent and cumulative performance) were more likely to prefer the best member over a select crowd, $\chi^2(1) = 6.65$, $p = .010$, a result apparent in Figure 7. Feedback did not affect the choice between the best member and the whole crowd, $\chi^2(1) = 1.69$, $p = .19$. In sum, information about the economists' performance encouraged participants to use a best-member strategy at the expense of using select crowds.

Performance. We used participants' MAEs over Trials 2–24 to evaluate performance, with lower scores indicating better performance. Across the 80 participants, MAEs ranged from 0.44 to 0.65 and averaged 0.51 ($SD = 0.05$). For reference, the 11 economists' MAEs ranged from 0.38 to 0.74 over these trials and averaged 0.55 ($SD = 0.11$). A policy of strictly averaging the whole crowd would have yielded an MAE of 0.48.

A 2 (choice) × 2 (feedback) between-subjects analysis of variance revealed only a main effect of choice on performance, $F(1, 76) = 5.32$, $p = .024$, $\eta_p^2 = .07$. On average, participants in the high-choice condition made lower forecast errors ($M = 0.50$, $SD = 0.04$) than those in the low-choice condition ($M = 0.53$, $SD = 0.05$). Neither the amount of feedback, $F(1, 76) = 0.95$, $p = .33$, nor its interaction with choice, $F(1, 76) = 0.28$, $p = .60$, affected performance.

In order to better understand the superior performance by those in the high-choice condition, we examined participants' AEs conditioned on their choice of strategy. Figure 8 plots performance on the *y*-axis against the number of economists selected on the *x*-axis for the low-choice and high-choice conditions (averaged over the feedback condition). As suggested by the figure, participants in the high-choice condition actually performed slightly worse than those in the low-choice condition on the two strategies they had in common (the best member and the whole crowd), $F(1, 71) = 4.13$, $p = .046$, $\eta_p^2 < .01$, though they differed markedly in how often they used these strategies (see Figure 7). However, because participants in the high-choice condition frequently chose to use select crowds, which had an MAE of 0.48 (across $k = 2$–7), they had lower forecast errors than participants in the low-choice condition.

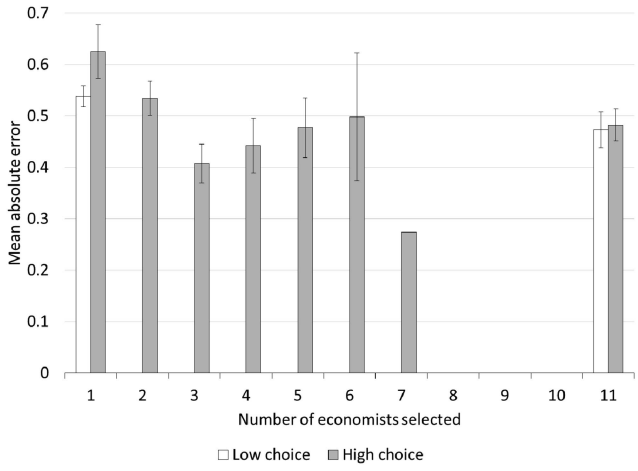

Figure 8. Actual performance conditioned on strategy in Experiment 2. Means with standard errors are reported. Lower values indicate better performance.

Discussion

When the option was available, people shifted from a best-member strategy to a select-crowd strategy, and their performance improved for it. Thus, the appeal of select crowds observed in Experiment 1 extended to people's behavior in this study on a task in which the strategy was not made particularly salient and participants were financially motivated to judge well. Nevertheless, the evidence also illustrated the strong allure of a best-member strategy. As seen in Figure 7, the more information participants had about the performance of the economists, the more likely they were to choose a single economist over a select crowd. When dispersion in expertise is high and past performance is diagnostic of true expertise, this tendency will serve people well. But it will backfire when dispersion in expertise is low or when past performance is an unreliable guide to expertise.

We conclude from these two experiments that select crowds are an attractive judgment strategy to people. In our final experiment, we look into the psychology behind this.

Experiment 3

We consider in our final experiment different explanations for people's attraction to select crowds. One hypothesis is that people believe averaging is more accurate than a best-member strategy. This seems unlikely as the main driver of the preference for select crowds, given prior research demonstrating a poor appreciation of the benefits of averaging (e.g., Larrick & Soll, 2006), as well as participants' general avoidance of the whole-crowd strategy in Experiments 1 and 2. To the extent people do endorse the accuracy of averaging, we expect them to use the whole-crowd strategy more and the best-member strategy less. The frequent selection of two to three economists in Experiment 2 suggests another hypothesis for the attraction of select crowds: The strategy allows people to hedge their uncertainty about the crowd's best member by selecting a small subset of the crowd's better members. This intuition is not wrong—uncertainty about the expert should push one to use select crowds or the whole crowd. But it is incomplete because it fails to recognize the role that bracketing plays in enhancing the performance of averaging, and that bracketing increases with more judges. We expect the more people believe in the unreliability of individual performance, the more they will prefer a select-crowd strategy to a best-member strategy.

Method

Sample. Two hundred and eleven adults from a national panel in the United States completed a short online study in exchange for

[7] The interaction between feedback and trial was not significant for either the low-choice condition, $\chi^2(1) = 0.02$, $p = .89$, or the high-choice condition, $\chi^2(2) = 5.01$, $p = .08$. We therefore focus only on the main effect of feedback controlling for trial.

Table 4
Questions Associated With the Choice of Committee Members in Experiment 3

No.	Question
A11	Averaging the forecasts of multiple committee members will be less accurate than choosing Amy to provide the forecast.[a]
A12	In general, the average forecast of multiple committee members will be closer to the truth than Amy's forecast.
A13	On balance, Amy's forecasts of attendance will be more accurate than the average forecast of multiple committee members.[a]
A21	Individual mistakes tend to cancel out when averaging the forecasts of multiple committee members.
A22	Averaging tends to amplify the mistakes of individual forecasters.[a]
A23	Individual biases are offset by averaging the forecasts of multiple committee members.
R11	Relying on Amy's forecast will lead to more variable results over time than averaging the forecasts of multiple committee members.
R12	The accuracy of Amy's forecast will fluctuate more wildly than the accuracy of the committee members' average forecast.
R13	Averaging will lead to smaller swings in accuracy from period to period than relying on Amy's forecast.
R21	It is easy to predict who will be the most accurate committee member in the future.[a]
R22	It is clear who the best forecaster will be in the future.[a]
R23	The past performance of each forecaster is a reliable guide to their future performance.[a]
S01	Choosing one committee member is a simpler strategy to use than averaging the forecasts of multiple committee members.[a]
S02	Averaging the forecasts of multiple committee members is more complicated than relying on the forecast of one member alone.[a]
S03	It is easier to go with one committee member than to average the forecasts of multiple committee members.[a]

Note. Answered on 7-point Likert scale anchored by *strongly disagree* and *strongly agree*. Composite scales were constructed so that higher scores favored averaging.
[a] Reverse-coded.

a small payment. All passed a one-question attention check to verify that they had read instructions before being assigned to their condition and viewing the stimuli. Eight participants provided incorrect answers to two or more questions (out of five) in the training phase of the exercise and were excluded from all analyses. This left a final sample of 203 participants (M_{age} = 36.2 years, 146 female).

Procedure. The experiment included a training phase and an evaluation phase. In the training phase, we introduced participants to the concept of forecast accuracy in the context of a college film committee that predicts attendance at upcoming events. To facilitate interpretation of how we measured expertise, MAE was described as the "Average Size of Miss" and carefully illustrated. We made clear to participants that smaller misses were preferred to larger misses when evaluating performance and that misses were sometimes over and sometimes under actual attendance. Participants answered five questions to test their understanding of the task. Those who incorrectly answered two or more of the questions were allowed to complete the study but were excluded from the analysis. (Full details of the training phase are omitted for brevity but are available from the authors.) In the evaluation phase, participants were presented with a table listing the five members of the film committee and their respective levels of expertise over 30 recent films ("Average Size of Miss"): Amy (20), Brad (25), Carrie (30), Doug (35), and Eric (40). Participants then completed two tasks, the order of which was counterbalanced. One task required participants to choose committee members:

> Imagine that you manage the theater on campus. You rely on the forecasts of the film committee for planning purposes. Whose forecasts would you prefer to include for predicting attendance at future films? You can include as many members as you like, including all five. If you include more than one member, the committee will provide the average forecast of those you selected. Select one or more committee member(s) to be included in your forecast.

The other task required participants to answer 15 questions addressing five variables relevant to the choice of committee members. These variables—two related to accuracy, two related to risk, and one related to ease of use—were identified in an earlier pilot study of 99 adults who read the same scenario, chose members of the committee, and listed reasons for their choices in an open-ended format. The variables and questions are presented in Table 4. Participants exited the study upon completion of the two tasks.

Results

Preliminary analysis. An exploratory factor analysis of participants' responses to the 15 questions identified five factors that adequately reproduced the observed covariance matrix, $\chi^2(40) = 44.04$, $p = .30$ (using maximum-likelihood estimation with oblique rotation). Questions A11, A12, and A13 were averaged for an indicator of the perceived accuracy of averaging multiple judgments (*averaging is more accurate*; $\alpha = .69$). A21 and A23 were averaged for an indicator of the perceived error-reducing benefits of averaging (*averaging reduces error*; $\alpha = .66$).[8] R11, R12, and R13 were averaged for an indicator of the perceived variability of averaging (*averaging is less variable*; $\alpha = .69$). R21, R22, and R23 were averaged for an indicator of the perceived unpredictability of individual performance (*performance is unpredictable*; $\alpha = .87$). And S01, S02, and S03 were averaged for an indicator of the perceived simplicity of averaging (*averaging is easier to use*; $\alpha = .82$). Higher scores on these measures indicate beliefs that favor averaging. Summary statistics are provided in Table 5.

Strategy. Across the two orders of task completion, 45 participants chose to rely on the forecast of only one committee member (22.2%), 82 chose to average the forecasts of two members (40.4%), 60 chose to average the forecasts of three members (29.6%), five chose to average the forecasts of four members (2.5%), and 11 chose to average the forecasts of all five members (5.4%). In our analyses, these choices were reduced to one of three basic strate-

[8] Including Question A22 in this composite reduced its reliability to .53, so it was dropped from consideration.

Table 5
Summary Statistics for Experiment 3

Variable	*M*	*SD*	Pearson correlations 1	2	3	4	5
1. Averaging is more accurate	3.48	1.39	—				
2. Averaging reduces error	5.09	1.06	.09	—			
3. Averaging is less variable	4.16	1.32	.32**	.26**	—		
4. Performance is unpredictable	3.89	1.52	.19**	−.06	.02	—	
5. Averaging is easier to use	3.63	1.46	.14*	−.26**	−.01	.31**	—

Note. $N = 203$.
* $p < .05$. ** $p < .01$.

gies: the best member (22.2%), select crowds (72.4%), and whole crowd (5.4%). The use of select crowds was marginally higher by those who completed their ratings before choosing ($p = .068$, Fisher's exact test).

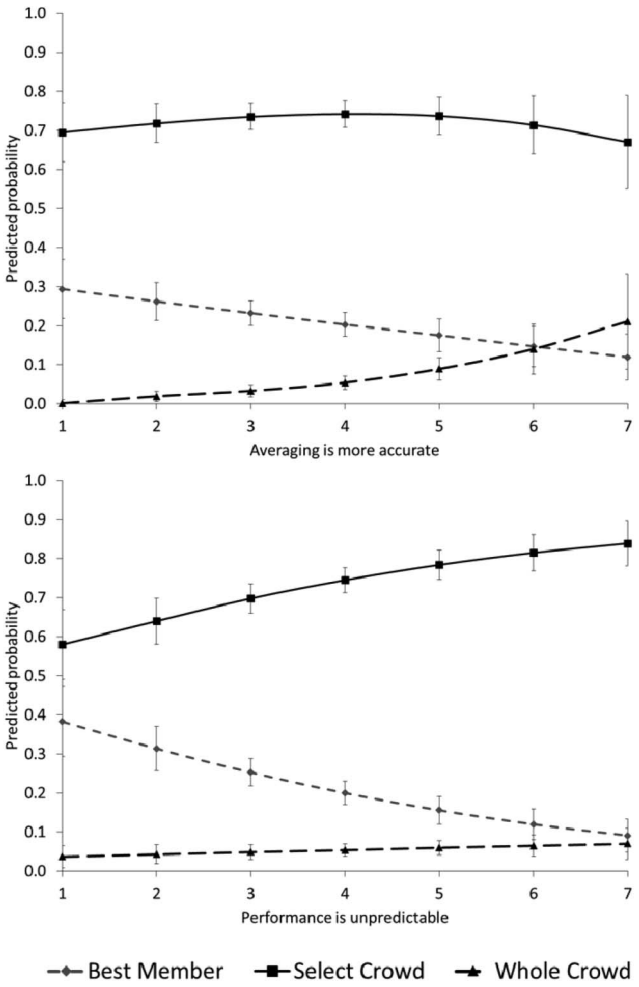

Figure 9. Predictors of strategy in Experiment 3. The likelihood of choosing the whole crowd increased with beliefs about the accuracy of averaging (top panel). The likelihood of choosing select crowds increased with beliefs about the unreliability of individual performance (bottom panel). Standard errors are shown.

To identify the factors correlated with these choices, we conducted a multinomial logistic regression of participants' chosen strategies on the five measured variables, controlling for task order. Use of the best-member strategy served as the baseline outcome in the model. The pseudo-R^2 was .088, $\chi^2(12) = 25.82$, $p = .011$, which allowed us to reject that there was no relationship between the predictors and choice of strategy. Two individual variables proved significant, and their effects are illustrated in Figure 9.[9] All else equal, favorable beliefs about the accuracy of averaging (*averaging is more accurate*) predicted greater use of the whole crowd relative to the best member ($B = 0.68$, $SE = 0.30$, $z = 2.28$, $p = .023$). However, these beliefs did not affect the relative use of select crowds ($B = 0.16$, $SE = 0.15$, $z = 1.06$, $p = .290$). In other words, participants who endorsed the accuracy of averaging were more likely to choose the whole crowd. The preference for select crowds over the best member, in contrast, was associated with beliefs about the predictability of individual performance. All else equal, participants who more strongly believed in the difficulty of identifying experts (*performance is unpredictable*) were more likely to choose select crowds instead of the best member ($B = 0.31$, $SE = 0.14$, $z = 2.24$, $p = .025$). These beliefs did not affect the relative use of the whole crowd ($B = 0.37$, $SE = 0.25$, $z = 1.47$, $p = .142$). In sum, participants who were dubious about the predictability of individual performance were more likely to choose select crowds.

Discussion

As shown in the first two experiments, people believe that select crowds are accurate, and they use select crowds when given the opportunity. In this third experiment, we explored more thoroughly the source of people's attraction to the select-crowd strategy. We found that there are two distinct beliefs that drive strategy choice. First, people tend to go with select crowds, at the expense of the best-member strategy, when they doubt the predictability of future performance. In the experiment, the greater the degree to which participants considered the past performance of the committee members a poor guide to their future performance, the more likely they were to select two to four committee members to assist with the forecast. Second, participants who endorsed the accuracy of averaging were more likely to use the whole crowd instead of

[9] Beliefs that *averaging reduces error*, $\chi^2(2) = 0.35$, $p = .84$; *averaging is less variable*, $\chi^2(2) = 2.15$, $p = .34$; and *averaging is easier to use*, $\chi^2(2) = 0.61$, $p = .74$, had no effects on the choice of strategy.

relying on one person. Our conclusion from these results is that the beliefs that guide people toward select crowds and whole crowds are different. People eschew the best member for select crowds when they think performance is hard to predict. In contrast, some people truly appreciate the wisdom of crowds, even if it is only on faith as opposed to a deep understanding of statistics (Larrick & Soll, 2006), and these individuals are more likely to average the whole crowd.

General Discussion

What is the best strategy for leveraging collective wisdom? There is a long tradition in social psychology recognizing the accuracy that can be achieved with statisticized groups (see, e.g., Stroop, 1932), which has gained currency as the wisdom of crowds (Surowiecki, 2004). However, outside of academic circles, there is considerable skepticism toward algorithmic approaches to judgment and decision making such as averaging, which leads people to search out the best member of a group instead. In fact, both strategies have true shortcomings. From a prescriptive standpoint, the best-member and whole-crowd strategies produce excellent outcomes in some judgment environments but poor outcomes in others. Our simulations and analyses of archival data demonstrated the strengths and limitations of each. The best-member strategy performs well when dispersion in expertise is high, which allows one to reliably identify the more capable members of a crowd, but poorly when dispersion is low. Averaging the whole crowd performs well when bracketing rates are high, which allows individual judgment errors to cancel out, but poorly when rates are low. The problem, of course, is that decision makers often have little information about the dimensions of their environment, which means they gamble on getting it right when choosing their judgment strategy and can lose if they guess wrong.

The psychology of judgment aggregation also limits the effectiveness of each strategy. People resist strategies that ignore their intuitions about expertise and give equal weight to all opinions. The growth of prediction markets, for example, which rely on the statistical aggregation of individual bets to forecast events, has been slow in part because "bosses may also be wary of relying on the judgments of non-experts. Yet many pilot projects run so far have shown that junior staff can often be surprisingly good forecasters" ("Prediction Markets: An Uncertain Future," 2009, p. 68). Similarly, the participants in our experiments had unfavorable opinions about the whole-crowd strategy and infrequently used it when paid to make accurate judgments. Instead, people are drawn to a best-member strategy, which performs admirably when there are large differences in ability and reliable cues for identifying the best member. But these conditions are often not met in practice, which means people will inevitably err in their choice of the best member. Moreover, people must often rely on imperfect cues, inferring that the most informed, confident, or talkative individual is the most capable member of a group (Anderson et al., 2012; Budescu et al., 2003; Littlepage et al., 1995; Sah et al., 2013).

Select crowds offer a solution to both types of shortcomings. Foremost, the performance of the strategy is more robust than that of the best-member or whole-crowd strategies. Whereas the performance of the two pure strategies is sensitive only to dispersion in expertise or to bracketing rates, select crowds respond to both (see Figure 3). As a consequence, select crowds outperform both the best member and the whole crowd in a significant number of representative judgment environments and are a close second in the remaining environments. Moreover, select crowds are less risky—across environments they have a much higher floor than the pure strategies (see Figure 5).

Select crowds are also psychologically attractive. When the option is salient, people are drawn to select crowds because they prefer to rely on experts but recognize that identifying the best member of a group is difficult. Select crowds allow people to indulge their strong preference for expertise but hedge their uncertainty about the best member by including additional opinions. Since people are better at choosing the better members of a crowd than choosing its best member, this should enable them to outperform both pure strategies. Although the use of select crowds in Experiment 2 was encouraging, participants would have done better with slightly larger crowds. This suggests that people's intuitive choices reflect an incomplete understanding of the strategy, which capitalizes on dispersion in expertise where it exists *and* on the error-cancellation benefits of bracketing that come with including more people.

We have argued that a select crowd of five judges effectively balances the benefits of relying on the crowd's more capable members and the bracketing that comes with averaging. Our rule of thumb, "take the top five," is easy, memorable, and effective (Heath & Heath, 2007). Select crowds of five performed well in the simulations and with empirical data. Even so, the results suggest that in some cases it might be helpful to include a few extra judges. For instance, Figure 5 indicates that eight to 10 judges may reduce downside risk compared with five judges (although at the expense of lower expected accuracy) and that for crowds of 100 people or more, there may be marginal gains in expected accuracy from adding a few judges. In general, however, most of the available benefits are obtained with the simple rule of using the top five judges.

In the remainder of this discussion, we address several other considerations. First, one might have information about the environment, and so it might be possible to do better than a select crowd with a strategy matched to the environment. Second, we consider two environments not covered by the simulations that pose a challenge for select crowds. Third, although we have focused here on historical performance as a cue to expertise, in principle the select-crowd strategy can be applied with any cue that correlates with individual ability. We consider the prospects of expressed confidence as such a cue. Finally, we discuss possible extensions of the select-crowd strategy beyond the quantitative estimation task considered here.

Knowledge of the Environment

The select-crowd strategy is appropriate when little is known about the environment because the strategy is robust across a wide range of situations. Sometimes, however, the decision maker may have useful information about the environment. In principle, a decision maker armed with such knowledge could outperform a select crowd by adapting the strategy to the environment. For instance, with information about dispersion in expertise and bracketing rates, a decision maker could average the whole crowd when dispersion is low and bracketing is high (see Figure 2A) and go with the best member when dispersion

is high and bracketing is low (see Figure 2D). Note that in both cases the chosen strategy has a relatively small margin of victory over the select crowd but that a wrong guess about the environment could lead it to perform much worse than the select-crowd strategy. For the adaptive strategy to be superior overall, therefore, one would need highly valid cues to the environmental parameters.

Decision makers may also be attracted to more sophisticated models when they have access to long histories. In this case, regression analysis can be run to calculate the optimal weights to place on individual judges. Regression captures an environment's parameters (dispersion in expertise and bracketing) in that it puts more weight on better judges and accounts for the correlation across answers from different judges. Relying on regression, however, has its limitations. In his foundational work on these techniques, Dawes (1979) distinguished between proper models that use regression analysis and improper models in which weights are chosen by some other means (e.g., averaging with equal weights). Proper models are by definition optimal when applied to the same data upon which they are built, but they tend to overfit the noise in these data and are less accurate when used to make predictions in a new sample of data. This can lead to some surprising results when data sets are not very large (Einhorn & Hogarth, 1975). In our context, a simple average of judges' opinions can be expected to perform as well as or even better than regression weights when there are fewer than 15–20 observations per judge (Dawes, 1979). Thus, for a crowd of 50 judges, one would need about 1,000 observations for a proper model to outperform simple averaging, based on this simple rule of thumb. Usually, this will be impractical. From this perspective, the select-crowd strategy is desirable because it is both practical and effective. It works because it combines the well-known robustness of averaging with the fact that in most contexts judges differ in their ability and select crowds take advantage of this.

There is a second reason why a select crowd might outperform a proper model such as regression, which is that the abilities or insights of the judges may change over time. For example, the industries that drive economic growth may change over time. This suggests that forecasters' relative abilities may change with general trends in the economy if they differ in their specialized knowledge about different industries. A select-crowd strategy that relies on recent history can adapt over time to include the forecasters whose knowledge best matches the current state of the world.

Variations on the Environment

Our simulations systematically varied dispersion in expertise and the degree of bracketing over wide ranges of parameter values. However, there are many other variations, and our analyses could not cover all of them. One type of variation has to do with the shape of the distribution. As mentioned in footnote 1, the basic result that select crowds are robust and often outperform both the best member and the whole crowd holds for different beta distributions, including approximations to a normal distribution and skewed distributions. However, there are some unique situations that are not captured well by a smooth and well-defined distribution. Below, we consider the case of a single expert amidst a crowd of novices as one example of this type of variation. A second type of variation we consider is when judges are clustered into types, such that judges belonging to the same cluster tend to produce more similar estimates compared to judges belonging to different clusters. An implication of this is that judges from the same cluster are less likely to bracket the truth.

A lone expert among novices. We envision here a situation in which a lone expert exists who is vastly superior to the rest of the crowd (cf. Davis-Stober et al., 2014). Our simulations modeled expertise as a continuous distribution. The case of a lone expert would be modeled as a discontinuous distribution with a very large difference in accuracy between the best member and the second best. An example could be a large, multifunctional team working on a project to design and sell a new smart phone. Although everyone may have an opinion, the marketing chief may be in the best position to provide a forecast of demand. A select crowd of five would include the expert, but its performance would be severely diluted by novices; it would be much better in this situation to just listen to the expert. In this environment, the best-member strategy is ideal, assuming that the expert can be reliably identified. However, we note two considerations. First, believing that one is in a lone-expert environment is not the same thing as actually being in one. Although a single crowd member may seem more knowledgeable than everyone else, research on group decision making has shown that people often guess wrong about who the best member is and that common cues to expertise are more fallible than one might think (Bonner, 2004; Budescu et al., 2003; Henry, 1995; Littlepage et al., 1995, 1997). Second, in many situations, the crowd is actually a subset of a much larger collective that has already been selected (or self-selected) for ability. For example, the 40 to 50 economists who participate regularly in the *Survey of Professional Forecasters* are an elite subset of the many thousands of people who might have an opinion about the trajectory of the economy. Such a crowd is likely to have relatively low dispersion in expertise and is unlikely to have only a single expert.

Clusters of judges. Experts may think alike, especially if they share training, information, or experience, or if they interact frequently (Larrick, Mannes, & Soll, 2011). Clusters of expertise may therefore develop within a crowd, with each cluster's members taking a common approach and providing similar estimates (Lamberson & Page, 2012). Our simulations did not encompass this possibility. Although judges differed in their ability in the simulations, there was no tendency for a particular subset of judges to be more alike in the direction of their errors when compared to judges outside the subset. The possibility of clustering poses a potential problem for the select-crowd strategy because members of a cluster will tend to give similar estimates and therefore receive similar ranks. A select crowd of five judges all arriving from the same cluster would lack the diversity of thought needed to produce bracketing. Consequently, averaging the top five would provide about the same accuracy benefits as relying on the best member.

One method for overcoming the clustering dilemma would be for the decision maker to avoid including multiple judges in the select crowd from the same cluster. Naturally, this would require cues to clustering. For example, one might notice that

several economists received their doctorates from the same university at about the same time. Alternatively, it might be noticed that several economists tend to give very similar forecasts, even when they are wrong. In each case, a point could be made to include at most only one member of the identified cluster in the select crowd. This approach entails a small risk. Accuracy could be worse than with the standard select crowd because it excludes some better, presumably redundant judges for the sake of diversity and, with hope, enhanced bracketing. We suspect that the risk is minimal, as long as the included judges still come from the upper portion of the ranking. Our reasoning is that even if dispersion in expertise is large within the whole crowd, it will be much less within a subset at the top (e.g., top 10 or top 15). The potential benefits from bracketing are likely to outweigh the potential harm of lowering the overall expertise in the crowd (Hogarth, 1978). Additional research is needed to establish whether or not people can beneficially fine-tune their select crowds to avoid redundancy and incorporate diversity.

Cues to Expertise

We focused on historical performance as the cue by which to rank judges and select the best. We showed that much can be accomplished with a history of around five past observations. This is because when there are large differences in expertise, it is easy to discriminate among judges and a short history is sufficient, and when there are small differences in expertise, there is not much to be gained by accurately ranking the judges.

In many situations, however, data on past performance may not be available, such as for unique events (e.g., how long a given foreign dictator will remain in power) or when record keeping has been poor. A decision maker may therefore look to other cues to ability to rank judges, such as their training, experience, or credentials. Here, we consider the validity of confidence as a cue to expertise (cf. Hertwig, 2012; Koriat, 2012). Recent reviews suggest that people's self-assessments are moderately valid indicators of their true ability (Ackerman, Beier, & Bowen, 2002; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Freund & Kasten, 2012). On judgment tasks like those presented in this research, experts and nonexperts tend to be overconfident in their knowledge (McKenzie, Liersch, & Yaniv, 2008; Moore & Healy, 2008; Tetlock, 2005). Therefore, it may be risky to follow the single most confident member of a group. Nevertheless, although confidence may not be diagnostic enough to justify choosing a best member, it may still be useful in identifying a suitable select crowd.

To explore this possibility, we analyzed 10 sets of our archival experimental data in which participants provided item-level confidence (on a 1–7 scale) in the accuracy of their point estimates on general knowledge questions (from Soll & Larrick, 2009, Experiment 4). For each question, we computed the correlation between participants' stated confidence and their accuracy (as measured by AE but reverse-scored so higher numbers indicate better performance). These correlations were averaged across questions for each set of data. The average correlation was .15, ranging from −.05 to .25, and the correlations were positive in nine of the 10 sets of data. For this sample, stated confidence was a modestly valid predictor of expertise, so we examined how well the select-crowd strategy would work if judges were chosen by their stated confidence.

Figure 10 presents the average results over 10,000 simulations of ranking and selecting judges by their stated confidence on each question. For this (small) sample of 10 data sets, selecting judges based on their confidence performed at a similar level as selecting judges based on their performance over five periods. Although selecting a single best member based solely on confidence is not a recipe for success, as illustrated in the figure, selecting a crowd of confident judges and averaging their opinions yields accurate judgments. This raises an interesting contrast between two uses of confidence. The most confident judges are almost always overconfident (Larrick et al., 2007; Moore & Healy, 2008; Tetlock, 2005) in that their statements of probability exceed the rate at which they are actually correct. For example, statements such as "I'm 90% sure that GNP will be between 5% and 6% next year" tend to be right only 50%–70% of the time (Moore & Healy, 2008; Soll & Klayman, 2004). Nevertheless, as long as confidence and actual knowledge are positively correlated (Ackerman et al., 2002; Freund & Kasten, 2012), the former can be used as a cue to select a crowd of forecasters. According to Figure 10, selecting the five most confident members of a crowd will yield a sample of forecasts that performs well when averaged together.

Extensions

Because not all tasks are easily reduced to numerical values, an important question is whether the select-crowd strategy generalizes to other tasks. Many forecasts are categorical, such as predicting whether a medical intervention will be successful, who will win the best-actress Oscar, or where to forage for food. These are essentially voting tasks. In an analysis of the foraging problem, Hastie and Kameda (2005) used simulation to compare a number of different strategies that groups could employ to choose among alternatives. They found that majority

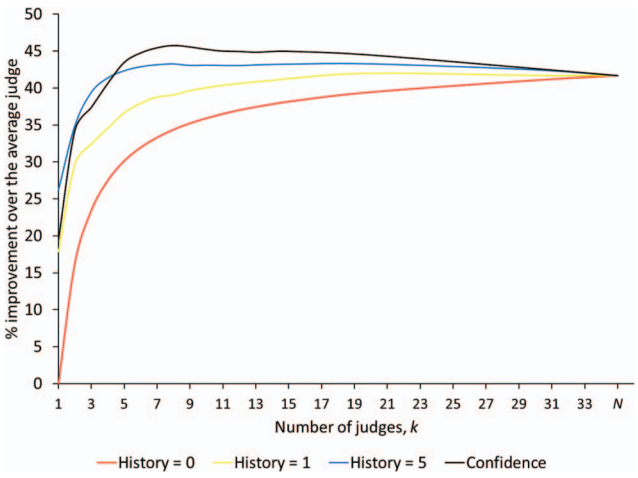

Figure 10. Performance of strategies in which judges are selected based on their stated confidence in 10 archival data sets. The performance of the best member is indicated at $k = 1$, of the whole crowd at $k = N$, and of select crowds at $1 < k < N$. The performance of judges selected based on histories of zero, one, and five trials is included for comparison.

(or plurality) rule was not only very simple but also one of the most effective ways to combine opinions. However, when dispersion in expertise was high, majority rule was supplanted by a best-member strategy, which parallels our result for quantity estimation. It would be interesting to explore whether select crowds are also a robust strategy in voting as in quantity estimation.

In many situations, categorical judgments are accompanied by a probability assessment (e.g., 70% chance of rain). Because outcomes are nominal (e.g., an outcome is coded as 1 if it rains and 0 otherwise), there is no obvious analogue to our bracketing measure, and performance is typically evaluated with specialized scoring rules, such as the Brier score. Nevertheless, recent work is suggestive that select crowds can perform as well as or better than more complex weighing procedures on this task (Budescu & Chen, in press).

Finally, outcomes that seem to be entirely subjective, such as people's taste in movies or consumer products, can also benefit from crowds. For subjective tasks, the truth is no longer an objective fact in the world (e.g., GDP in the third quarter) but an individual's own preferences. Websites such as Rotten Tomatoes and Metacritic, for instance, provide summaries of crowds' ratings for movies and television shows. Other rating systems, such as Wine Spectator's, rely on one expert's judgment. A popular and well-researched alternative to these two approaches is collaborative filtering, a method in which the tastes of one individual are predicted from the tastes of similar others. Although the mechanics of collaborative filtering can be complex, it is analogous to the select-crowd strategy in that the preferences of one individual are predicted based on a select crowd of others who share the person's tastes, interests, or preferences based on a history of previous choices.

Conclusion

Research on benefiting from the opinions of others has traditionally focused on strategies that emphasize individual expertise or those that emphasize collective wisdom. We have argued that this is a false dilemma. Select crowds take advantage of expertise where it exists without sacrificing the reduction in error associated with aggregating the judgments of a large and diverse crowd. It is therefore a robust strategy for improving judgment across a wide array of environments. And because it is robust, the strategy frees the decision maker from having to identify whether the environment favors a best-member strategy or averaging the whole crowd (and from bearing the costs of guessing wrong). It is also has intuitive appeal to decision makers, who are generally skeptical about the wisdom of crowds but recognize the fallibility of identifying the crowd's best member. Select crowds blend the power of statisticized groups, which have a rich tradition in social psychology, with the intuitive allure of finding and relying on experts.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33,* 587–605. doi:10.1016/S0191-8869(01)00174-X

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology, 103,* 718–735. doi:10.1037/a0029395

Armstrong, J. S. (Ed.). (2001). *Principles of forecasting: A handbook for researchers and practitioners*. doi:10.1007/978-0-306-47630-3

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–804. doi:10.1037/0022-3514.45.4.792

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

Bonner, B. L. (2004). Expertise in group problem solving: Recognition, social combination, and performance. *Group Dynamics: Theory, Research, and Practice, 8,* 277–290. doi:10.1037/1089-2699.8.4.277

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes, 88,* 719–736. doi:10.1016/S0749-5978(02)00010-9

Bottger, P. C. (1984). Expertise and air time as bases of actual and perceived influence in problem-solving groups. *Journal of Applied Psychology, 69,* 214–221. doi:10.1037/0021-9010.69.2.214

Budescu, D. V., & Chen, E. (in press). Identifying expertise and using it to extract the wisdom of crowds. *Management Science.*

Budescu, D. V., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes, 90,* 178–194. doi:10.1016/S0749-5978(02)00516-2

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5,* 559–583. doi:10.1016/0169-2070(89)90012-5

Csaszar, F. A., & Eggers, J. P. (2013). Organizational decision making: An information aggregation view. *Management Science, 59,* 2257–2277. doi:10.1287/mnsc.1120.1698

Davis, J. H. (1996). Group decision making and quantitative judgments: A consensus model. In E. H. Witte & J. H. Davis (Eds.), *Understanding group behavior: Consensual action by small groups* (pp. 35–59). Mahwah, NJ: Erlbaum.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*. Advance online publication. doi:10.1037/dec0000004

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

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology, 51,* 629–636. doi:10.1037/h0046408

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87. doi:10.1111/1467-8721.01235

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171–192. doi:10.1016/0030-5073(75)90044-6

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84,* 158–172. doi:10.1037/0033-2909.84.1.158

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting, 15,* 227–246. doi:10.1016/S0169-2070(98)00073-9

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138,* 296–321. doi:10.1037/a0026556

Galton, F. (1907, March 28). The ballot-box. *Nature, 75,* 509–510. doi:10.1038/075509f0

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121,* 149–167. doi:10.1037/0033-2909.121.1.149

Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315–342). Englewood Cliffs, NJ: Prentice-Hall.

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

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 129–157). London, England: JAI Press.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review, 112,* 494–508. doi:10.1037/0033-295X.112.2.494

Heath, C., & Heath, D. (2007). *Made to stick: Why some ideas survive and others die*. New York, NY: Random House.

Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes, 62,* 190–197. doi:10.1006/obhd.1995.1042

Hertwig, R. (2012, April 20). Tapping into the wisdom of the crowd—with confidence. *Science, 336,* 303–304. doi:10.1126/science.1221403

Hill, G. W. (1982). Group versus individual performance: Are $N + 1$ heads better than one? *Psychological Bulletin, 91,* 517–539. doi:10.1037/0033-2909.91.3.517

Hinsz, V. B. (1999). Group decision making with responses of a quantitative nature: The theory of social decision schemes for quantities. *Organizational Behavior and Human Decision Processes, 80,* 28–49. doi:10.1006/obhd.1999.2853

Hogarth, R. M. (1978). Note on aggregating opinions. *Organizational Behavior and Human Performance, 21,* 40–46. doi:10.1016/0030-5073(78)90037-5

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103,* 687–719. doi:10.1037/0033-295X.103.4.687

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55,* 623–655. doi:10.1146/annurev.psych.55.090902.142009

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting? A social psychological analysis. *International Journal of Forecasting, 27,* 14–40. doi:10.1016/j.ijforecast.2010.02.001

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310. doi:10.1037/0033-2909.107.3.296

Koriat, A. (2012, April 20). When are two heads better than one and why? *Science, 336,* 360–362. doi:10.1126/science.1216549

Lamberson, P. J., & Page, S. E. (2012). Optimal forecasting groups. *Management Science, 58,* 805–810. doi:10.1287/mnsc.1110.1441

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes, 102,* 76–94. doi:10.1016/j.obhdp.2006.10.002

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2011). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *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

Libby, R., Trotman, K. T., & Zimmer, I. (1987). Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology, 72,* 81–87. doi:10.1037/0021-9010.72.1.81

Littlepage, G. E., Robison, W., & Reddington, K. (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes, 69,* 133–147. doi:10.1006/obhd.1997.2677

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & Frost, A. G. (1995). An input–process–output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology, 69,* 877–889. doi:10.1037/0022-3514.69.5.877

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychological Bulletin, 55,* 337–372. doi:10.1037/h0042344

Mannes, A. E. (2009). Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Management Science, 55,* 1267–1279. doi:10.1287/mnsc.1090.1031

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes, 107,* 179–191. doi:10.1016/j.obhdp.2008.02.007

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*. Advance online publication. doi:10.1177/0956797614524255

Miner, F. C., Jr. (1984). Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance, 33,* 112–124. doi:10.1016/0030-5073(84)90014-X

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 502–517. doi:10.1037/0033-295X.115.2.502

Moore, D. A., & Klein, W. M. P. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes, 107,* 60–74. doi:10.1016/j.obhdp.2008.02.005

Moore, D. A., & Small, D. (2008). When it is rational for the majority to believe that they are better than average. In J. I. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 141–174). New York, NY: Psychology Press.

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

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552. doi:10.1037/0278-7393.14.3.534

Plato. (2005). *Essential dialogues of Plato* (B. Jowett, Trans.). New York, NY: Barnes & Noble.

Prediction markets: An uncertain future. (2009, February 28). *The Economist, 390,* 68.

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. London, England: Printer & Martin.

Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes, 121,* 246–255. doi:10.1016/j.obhdp.2013.02.001

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 299–314. doi:10.1037/0278-7393.30.2.299

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

Steiner, I. D. (1972). *Group process and productivity*. New York, NY: Academic Press.

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology, 15,* 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*. London, England: Little, Brown.

Tetlock, P. E. (2005). *Expert political judgment*. Princeton, NJ: Princeton University Press.

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. doi:10.1126/science.185.4157.1124

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243–268. doi:10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science, 13,* 75–78. doi:10.1111/j.0963-7214.2004.00278.x

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

Yetton, P. W., & Bottger, P. C. (1982). Individual versus group problem solving: An empirical test of a best-member strategy. *Organizational Behavior and Human Performance, 29,* 307–321. doi:10.1016/0030-5073(82)90248-3

Appendix

Procedure for Calculating the Performance of the Best Member, Whole Crowd, and Select Crowds in the Archival Data Sets

The procedure for calculating the performance of the best member, whole crowd, and select crowds was similar across the experimental and economic data, with a few exceptions. (From the experimental and economic data, respectively, we use the terms *estimate* and *forecast* interchangeably and *question* and *period* interchangeably.)

1. Calculate the performance of the average judge. We calculated the accuracy of each estimate by comparing it to the correct answer for the experimental data or to the realized value of the indicator for the economic data. For all items except nominal gross domestic product (NGDP) in the economic data, we used absolute error (AE) to measure accuracy, as in the simulations. (Because the scale of NGDP changes over time, we used absolute percentage error [APE] to measure accuracy for it instead of AE.) The mean of these AEs (or APEs) first within question (i.e., across judges) and then across questions indexed the performance of the average judge for each set of data.

2. Calculate the performance of the whole crowd. For each question, we took the average of the judges' estimates and compared it to the correct answer or to the realized value of the economic indicator.[A1] We calculated the AE of the judges' forecasts (or APE in the case of NGDP) and averaged these across questions to index the performance of averaging the whole crowd for each set of data.

3. Rank the judges. For each question, the judges were ranked based on their historical performance (i.e., on their AEs or APEs). As in the simulations, we considered multiple levels of history. For one period of history ($h = 1$), the judges were ranked on their performance in the prior period only. For two to five periods of history, they were ranked on their mean performance over the prior two to five periods (equally weighted). We also ranked and selected the judges at random ($h = 0$).

[A1] Our results did not materially differ if the median replaced the mean as the average judgment for whole crowds and select crowds. As in the simulations (see footnote 3 in the main text), median judgments were slightly more accurate, but this difference was small for select crowds.

We clarify three aspects of the ranking procedure. First, because the economic data were a time series, ranking judges on their past history of performance was a logical and straightforward procedure: A judge's rank in time t was determined by his or her AE (for $h = 1$) or MAE (for $h > 1$) in time $t - 1$. The experimental data, however, were not a time series, and participants may have answered the questions in a different order. A participant's rank for question q based on his or her AE (for $h = 1$) or MAE (for $h > 1$) for question $q - 1$ presumes people answered the questions in that order, which may have not been the case. To address this, we randomly ordered the questions at the start of each simulation and ranked participants accordingly. Second, in the economic data, judges frequently entered, exited, and reentered the survey. We chose not to require a full set of relevant history for a judge to be included in the rankings, primarily because requiring full participation would have dramatically reduced the number of eligible economists for longer histories. So, for example, any ranking based on a history of five periods included any economist participating in one to five of the prior periods. Third, in the economic data, rankings were based on like horizons. That is, when ranking performance for a given forecast horizon, we only considered prior performance for the same horizon. For example, when ranking economists for the current quarter (Horizon 2), we looked at their past performance making Horizon 2 forecasts.

4. Calculate the performance of the best member. After ranking the judges, we selected for each level of history the top-ranked judge ($k = 1$) and calculated the AE (or APE) of his or her estimate for the current period. The mean of these AEs (or APEs) across periods indexed the performance of the best member for each set of data and level of history.

Note that judges frequently tied for the top position. These ties were broken at random. Moreover, in the economic data only, the forecast of the panel's best member was at times not available because the judge did not participate in the current survey. On these occasions the next ranked judge who participated in the survey was selected as the best member. For example, if the economist ranked first in the last period did not participate in the current survey, the economist ranked second was selected as the current period's best member.

5. Calculate the performance of select crowds. After ranking the judges, we selected for each level of history the top k judges and calculated the AE (or APE) of their average estimate for the current period. (In the experimental data, we evaluated $k = 2$–12 for the 20 sets of estimates with $N \leq 20$ judges and $k = 2$–15, 20, and 30 for the 20 sets with $N > 20$ judges. In the economic data, we evaluated $k = 2$–15 and 20 for all 50 sets of forecasts.) The mean of these AEs (or APEs) across periods indexed the performance of select crowds for each set of data, level of history, and k. As in Step 4, ties among judges for a ranked position were broken at random. In the economic data, if a ranked economist did not participate in the survey, the next ranked economist was selected in his or her place.

6. Rerun the calculations. Sampling error is introduced in Steps 1–5 (a) by question order in the experimental data, (b) when randomly ranking judges for a history of 0, and (c) when tied ranks are broken at random. To reduce the influence of sampling error, we repeated Steps 1–5 10,000 times for the experimental data, 3,000 times for a history of zero in the economic data, and 100 times for histories of one to five in the economic data. We averaged the performance of the judgment strategies over these iterations.

7. Scale performance. To compare results across the sets of forecasts, we scaled the average performance of the best member, the whole crowd, and select crowds from Step 6 against the performance of the average judge from Step 1. Specifically, we calculated the percent improvement of the strategy over the performance of the average judge for each set of data. For example, the average error of a two-person select crowd's average forecast of the consumer price index (Horizon 2, history of two periods) was 0.95. The average economist's error was 1.11. Thus, the two-person select crowd outperformed the average economist by 14.4%. Higher percent improvements indicate better performance.

Received September 21, 2013
Revision received February 12, 2014
Accepted February 27, 2014 ■