The Wisdom of Small Crowds

Albert E. Mannes

University of Pennsylvania

Jack B. Soll and Richard P. Larrick

Duke University

Manuscript in preparation

Author Note

Correspondence concerning this article should be addressed to Albert E. Mannes, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall 517, 3730 Walnut Street, Philadelphia, PA 19104. Email: albert.mannes@gmail.com

Abstract

We present the results of four studies that introduce the virtues of small crowds, which are well-chosen groups of 3 to 6 experts. In Study 1, we demonstrate with experimental and archival data that the average judgment of a small crowd is more accurate than the judgment of best members with near certainty and regularly more accurate than the average judgment of much larger crowds. We next illustrate the intuitive appeal of small crowds to people in both their beliefs (Study 2) and their behavior (Study 3). In Study 4 we show that this preference for small crowds is related to people's beliefs about the reliability of individual performance. When the small-crowd option is made salient, people continue to "chase expertise" but hedge their ability to identify best members by choosing at least two experts.

The Wisdom of Small Crowds

The Chief Financial Officer (CFO) of a small but growing U.S. manufacturing company opens her *Wall Street Journal* on the morning of January 3, 2006. There she finds the six-month forecasts for several economic indicators made by a distinguished panel of roughly 50 macroeconomists. She also learns that one economist in particular—Sung Won Sohn—was the most accurate forecaster of these indicators over the prior six months (Gerena-Morales & Hilsenrath, 2006). Because this company bases its investment decisions largely on her advice about future economic conditions, the CFO considers her choice thoughtfully: Does she rely on the forecasts of the panel's ostensible expert, Mr. Sohn, or on the aggregate judgment of the panel in some form?

In this paper we address the CFO's response from two perspectives. We first discuss her decision from a normative standpoint. How should one confronted with a set of diverse and perhaps conflicting opinions use them in order to make the best judgment possible? Although a range of strategies are available, we contrast the performance of two alternatives typically studied in the psychology and forecasting literatures: relying on the panel's most accurate member, or averaging the forecasts of the entire panel. Next, we discuss people's typical behavior in these situations and dissect the psychology behind it. Empirical evidence on the whole indicates that people are reluctant to use averaged judgments and prefer instead to rely on experts. We then introduce the "small-crowd" strategy as an alternative to either relying on an expert or averaging the judgments of an entire panel. In four studies we demonstrate that this strategy (1) performs as well if not better than averaging the entire crowd under a wide range of conditions, and (2) is intuitively attractive and used as a judgment strategy when offered to

people. Relying on the wisdom of small crowds thus represents a sensible balance of both normative and descriptive considerations as a technique for improving human judgment.

Best Members and Averaging as Judgment Strategies

Research on how best to use the opinions of others in decision making has been characterized by two dominant themes. One is grounded in the literature on group decision making and emphasizes the importance of correctly weighting members' contributions commensurate with their expertise (Hackman, 1987; Libby, Trotman, & Zimmer, 1987; Yetton & Bottger, 1982). In the extreme, groups should rely entirely on its most expert member in order to avoid any process loss (Steiner, 1972). Socrates himself was perhaps the first advocate of this *best-member* strategy:

> "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).

The best-member strategy can easily be extended beyond the group context to any situation in which a decision maker is confronted with a panel of opinions. To judge well, the decision maker should identify the most qualified member of the panel and rely on his or her opinion.

In practice, however, identifying the crowd's best member is difficult. For one, it requires reliable and valid cues to expertise, such as a history of diagnostic performance (Baumann & Bonner, 2004; Littlepage, Robison, & Reddington, 1997) or diagnostic credentials, which are costly to collect and may be hard to come by in probabilistic, unpredictable environments. Thus despite Mr. Sohn's superior forecasting performance in the second half of

2005, his forecasts of GDP growth and inflation for the first half of 2006 ranked 34th and 32nd in accuracy, respectively. Identifying the best member also requires sufficient variability in performance among people to meaningfully designate one as truly the best, which will not be the case if the task is either very easy or very hard (Baumann & Bonner, 2004; Bonner, Baumann, & Dalal, 2002; Libby et al., 1987). Finally, the ability to interact with individual members of a group, panel, or crowd is no panacea. Although experts may be able to demonstrate the truth of their opinions for purely intellective judgments (Laughlin & Ellis, 1986), these are arguably rare in organizational settings. Instead people are often misled by non-diagnostic cues to expertise, such as confidence, talkativeness, or information differences (Budescu, Rantilla, Yu, & Karelitz, 2003; Littlepage, Schmidt, Whisler, & Frost, 1995; Sniezek & Van Swol, 2001).

In clear distinction to relying on best members, a second research stream has emphasized an alternative strategy: exploiting the wisdom of crowds (Surowiecki, 2004). This was famously demonstrated by Francis Galton (1907), who reported that the dressed weight of an ox on display at the local fair was only nine pounds less than the median estimate of nearly 800 spectators. Since then, numerous studies have demonstrated that simple heuristics for aggregating group judgments (such as the median or mean for numerical judgments or majority vote for categorical ones) perform as well or better than more complex strategies (for reviews, see Bates & Granger, 1969; Clemen, 1989; Gigone & Hastie, 1997; Hastie, 1986; Hastie & Kameda, 2005; Hill, 1982; Lorge, Fox, Davitz, & Brenner, 1958). For instance, the arithmetic mean of a panel's forecasts will, except in rare circumstances, prove superior in quality to the forecast of the panel's average member. For numerical judgments, that benefit arises through an error-correction process in which optimistic assessments by some panel members are offset by the pessimistic assessments of others. Hogarth (1978) demonstrated that the performance of averaging depends on two

factors: the level of expertise in the panel and the independence of its members (as indicated by the correlations among their judgment errors). The beauty of averaging is that panels with less expert but *independent* judges can outperform smarter panels of non-independent judges. Because the opinions of the first group will be more dissimilar, they are more likely to "bracket" the truth (Larrick & Soll, 2006), which is necessary for the correction of individual errors.

The choice between a best-member strategy and averaging depends on the environment (Payne, Bettman, & Johnson, 1988; Soll & Larrick, 2009). Clearly there are circumstances where a best-member strategy 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. In general, however, averaging is preferred when there is low to moderate dispersion in expertise, moderate to high bracketing rates (the degree to which the truth falls within the range of opinions), and low to moderate probability of correctly identifying the best member. Accordingly, averaging performs well across a wide range of environments (Einhorn, Hogarth, & Klempner, 1977; Soll & Larrick, 2009).

Although typically treated separately, both the best-member and averaging strategies are special cases of a general linear model in which individual judgments are weighted and additively combined. With a best-member strategy, the decision maker places a unit weight on the forecast of the panel member judged most expert and weights of zero on the forecasts of all other members. This makes it akin to a take-the-best heuristic (Gigerenzer & Goldstein, 1999) or a strict lexicographic rule (Payne et al., 1988) in which panel members are ranked based on one attribute or cue—in this case, their expertise—and the highest ranked member determines the decision. With simple averaging, the decision maker places equal weight on each member's

forecast (i.e., 1/*N*). In the studies presented herein, we compare best members and averaging—two pure strategies—with small crowds, a hybrid strategy which incorporates elements of both.

Chasing Expertise

How people actually incorporate others' opinions into their own beliefs has received empirical attention from several traditions. Social psychologists, for one, have documented the powerful normative influence groups exert on individuals to conform to a group answer (e.g., Asch, 1951). These studies typically focus on how individuals express public judgments in groups, and so a primary influence on their judgment is the desire to be accepted by others. We are instead principally concerned with how people choose to use the judgments of others when their primary motivation is judgment accuracy. In other words, we focus on the "informational influence" of outside sources (Deutsch & Gerard, 1955; Festinger, 1954; Kaplan & Miller, 1987).

Because judgment policies are influenced by whether the decision maker has formed an opinion independently of the group (Harvey & Harries, 2004; Soll & Mannes, 2011), we review two streams of research. First are the findings from studies of advice-taking. In this paradigm, individuals 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). The typical finding is that people egocentrically weight their own opinions—they place more weight on their initial beliefs than is normatively justified (Harvey & Harries, 2004; Mannes, 2009; Soll & Larrick, 2009; Soll & Mannes, 2011; Yaniv & Kleinberger, 2000; Yaniv & Milyavsky, 2007). For example, people receiving advice from another person about numerical outcomes tend to "adjust" their final beliefs about 30% towards the advice rather than taking a simple average (an adjustment of 50%). Yet this fails to tell the whole story. Rather than take a

simple or weighted average of their own opinion and the advice, people often “chase the expert” (Larrick & Soll, 2006). That is, people quite frequently choose one opinion over the other(s) depending on which they believe is closer to the truth. In one representative study, people chose their own opinion 36.1% of the time, chose the advisor’s opinion 10.0% of the time, and averaged (adjustments of 40–60%) only 17.5% of the time (Soll & Larrick, 2009). These choices, moreover, are driven by perceived differences in expertise, which often erroneously favor the self (Soll & Mannes, 2011).

Research in which the decision maker lacks a prior opinion reveals a similar but less pronounced pattern in judgment aggregation. In the panel-of-experts paradigm, individuals must form an opinion based on information provided by others. In this situation, judgment strategies are best described by a weighted-average model in which the weights correspond to the perceived expertise of the panel members (Birnbaum & Mellers, 1983; Budescu et al., 2003; Fischer & Harvey, 1999; Maines, 1996; Soll & Mannes, 2011). Again, however, weighted-averaging at the aggregate level can be misleading. For example, Fischer and Harvey (1999) demonstrated that absent feedback about the expertise of four forecasters, participants placed equal weight on each (as indicated by their standardized regression coefficients), which implies simple averaging. Across trials, however, a similar conclusion could be reached if a participant chose each forecaster exclusively 25% percent of the time. Soll and Mannes (2011) looked at trial-level weights and found that simple averaging (defined as adjustments of 40–60%) was common when the judges were equally diagnostic. When the judges differed in expertise, simple averaging declined while choosing became more prevalent.

The overall tendency that emerges from this research is that people chase expertise. At the aggregate level, this often appears weighted-averaging with weights that correspond to

perceived expertise. But if one person appears more knowledgeable than the others, he or she is often selected exclusively—the best-member strategy. Only in rare circumstances, when there is a paucity of cues to expertise, will people take simple averages.

Why is this case? One reason is that people do not appreciate the value of averaging (Larrick & Soll, 2006). The flawed intuitions of many are that averaging "locks in mediocrity" by guaranteeing the performance of the average judge. We asked 96 adults (M_{age} = 35.5 years, 46 female, 77 Caucasian) from a national online panel to evaluate methods of forecasting attendance at a hypothetical film series. Participants were introduced to five members of a committee for this film series who varied in their ability to accurately forecast attendance at upcoming events. Namely, the best forecaster erred by an average of 10 people per film, the second best by 20 people, the third best by 30 people, the fourth best by 40 people, and the worst forecaster by 50 people. When asked for their best estimate of the error they expected from averaging the forecasts of all five committee members for each film, 67 participants responded with 30 people. In other words, a majority of the participants believed averaging would perform at the average level of the committee.

A second reason people chase expertise is overconfidence in their ability to identify best members. Although some studies have found positive associations between perceived and actual expertise (e.g., Bonner et al., 2002; Bottger, 1984; Littlepage et al., 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 et al., 1995; Miner, 1984). A fair compromise may be that people are reliably capable of identifying the better members of a group but not its best member (Bonner, 2004; Miner, 1984). In practice people are often led astray by imperfect cues to expertise such as expressed confidence (Sniezek & Van Swol, 2001),

talkativeness (Littlepage et al., 1995), or how much information a judge possesses (Budescu et al., 2003).

The Wisdom of Small Crowds

We introduce the small-crowd strategy as a device which acknowledges the robust behavioral tendency to chase expertise yet still allows decision makers to exploit the mathematical benefits of averaging multiple opinions. The strategy involves ranking members of the group, panel, or crowd on the basis of expertise and then taking a simple average of the judgments made by its top k members, where $1 < k < N$ but is typically three to six judges.[1] As such, the small-crowd strategy falls within the general linear model discussed earlier. (In this case, weights of $1/k$ are placed on the judgments of the selected members and weights of zero are placed on all others.)

The small-crowd strategy has advantages over best members and averaging. Compared to best members, small crowds preserve the ability for the errors or biases of one judge to be offset by those of another. And while both best members and small crowds require assessments of expertise, small crowds are more forgiving—the best person in the group need not be identified with certainty, only the better members. Compared to averaging, small crowds exclude the least knowledgeable members of the group. This has two benefits. First, it should enhance the attractiveness and use of small crowds. In the aforementioned film-series study, we also asked participants to estimate the expected error of averaging the forecasts of just the two best forecasters on the committee (who, again, erred on average by 10 and 20 people). Fifty-nine of the 96 participants responded with 15 people, which was equivalent to the average

[1] Selection of $k = 1$ and $k = N$ judges are equivalent to best members and averaging, respectively.

performance of this 2-person small crowd. Thus we expect people to prefer "locking in" an error of 15 with a small crowd than an error of 30 with averaging.

Second, by excluding the worst performing members of a group, small crowds should perform as well as or better than averaging everyone. Note that improvements in accuracy from averaging exhibit diminishing returns with increases in group size (Hogarth, 1978). Libby and Blashfield (1978) suggested that the majority of averaging's benefits accrue with the just the first three judges. The small-crowd strategy performs well because it takes advantage of real expertise when it exists, and it takes advantage of averaging's error-correction benefits when differences in expertise are hard to identify. If there are valid cues to expertise, the small-crowd strategy selects a sample of experts whose mean judgment can outperform the average of the whole crowd. Even when a best member is known with certainty, the error-correction process of averaging will often outperform the best member. When there are no reliable differences in expertise or cues to expertise are weak, the small crowd strategy has two advantages. First, the small crowd represents a hedge against error in identifying the best member by creating a portfolio of "better" members on average. Second, even with abysmal cues that allow no detection of expertise differences, the small crowd strategy offers the error correction process of averaging multiple judges—where, once again, the first three to five judges offer the majority of benefits.

These benefits come with a price, however. Unlike averaging, the small-crowd strategy requires cues to expertise. Ideally, past performance of the panel members is both available and predictive of future performance. Often, though, that will be hard to come by, and decision makers will be left to rely on less diagnostic information. This risks selection of a panel's poorer rather than better members. Moreover, it is costly to collect this data, particularly if multiple

periods of performance are needed to reliably predict future performance. Finally, it is difficult to specify ex ante the optimal value of *k*. How robust, for instance, are 3-person small crowds across environments and tasks? To address this, we test a range of simple rules, such as "choose 3" or "choose 10".

Overview of Studies

We examine these and other considerations in four studies. Study 1 is an empirical analysis of the performance of the small crowds in two sets of data: numerical estimates made by experimental participants in laboratory settings, and real-world forecasts of various economic indicators made by panels of economists. We find that a small crowd of three to six judges consistently outperforms best members and performs as well if not better than averaging the entire panel. Study 2 examines people's intuitive beliefs about the wisdom of small crowds, and Study 3 compares the relative use of best members, small crowds, and averaging in an experiment with incentive-compatible conditions. We find a clear preference for chasing the expert over averaging, but we also find that people like and use small crowds when the option is available. Study 4 identifies beliefs about the accuracy of averaging and the reliability of individual performance as predictors of people's choices.

Study 1

Method

Materials. We assembled two types of data. The *experimental data* comprised 40 datasets featuring nearly 1,400 participants and 31,700 numerical estimates made in six laboratory studies published since 2007 (Larrick, Burson, & Soll, 2007; Mannes, 2009; Moore & Klein, 2008; Moore & Small, 2008; Soll & Larrick, 2009; Soll & Mannes, 2011). Each study recorded the private and independent responses of participants to questions from one or multiple

domains (e.g., temperatures, distances, prices, weights, etc.). Only participants with complete sets of responses were included. Eighteen of the 40 datasets included financial incentives for accurate estimates.

The *economic data* comprised approximately 28,000 forecasts made for six U.S. economic indicators by professional economists from 1969 to 2009. Forecasts were provided by the quarterly *Survey of Professional Forecasters* and actual, realized values of the indicators by the *Real-Time Data Set for Macroeconomists*, both publicly available at the Federal Reserve Bank of Philadelphia's website (http://www.phil.frb.org/index.cfm). The number of forecasts, periods covered, and forecasters differed by indicator and are summarized in Table 1. All forecasts were for six months from the date of the survey (shorter and longer forecast periods, though available, were not considered in this study).

Strategies. We compared the accuracy of estimates created using the following judgment strategies. *Averaging* used the arithmetic mean of all estimates (or forecasts) for that question or period (henceforth referred to as *trials*). *Best members* required ranking the judges based on prior or cumulative performance and using the current estimate of the top-ranked judge, with tied rankings broken at random.[2] *Small crowds* required ranking the judges based on prior

[2] For the experimental data, cumulative performance weighted all prior periods equally. For the economic data, which were time series, cumulative performance was based on simple exponential smoothing ($\alpha = .30$) so that performance in recent periods influenced a judge's ranking more than performance in distant periods. Note that random judges were always selected in the first period because prior performance data was absent.

or cumulative performance and using the mean estimate of the top *k* judges, with tied rankings broken at random.[3] Small crowds ranged from two to 10 people.

Additional strategies were included for comparison. The *trimmed average* involved ordering the estimates of all judges in a period, identifying the 10^{th} and 90^{th} percentiles of the distribution, and averaging the 80% of estimates (approximately) falling within that interval. The *median* involved using the median of all judges' estimates in a period. Finally, we included the estimate or mean estimate of, respectively, best members and small crowds selected at *random* each trial.

Procedure. We calculated the absolute error of each estimate (or mean estimate) by comparing it to the true value of the criterion. These absolute errors were normed to allow aggregation across datasets and time by dividing them by the average absolute error of the panel members that trial. As an illustration, consider estimates of temperature by a panel of judges, who on average erred in their guesses by 12 degrees. If the absolute error of the panel's average estimate was 8, and the absolute error of the panel's best member was 6, the normed values for the averaging and best member strategies would be, respectively, .67 and .50, with lower values indicating superior performance.

[3] Because breaking ties at random introduces sampling error, the process of ranking and selecting judges was repeated 100 times for each strategy. Mean performance was then calculated over these simulations. Moreover, it was common for forecasters in the economic datasets to exit the panel after a period either permanently or temporarily. Accordingly, for the best-member and small-crowd strategies, the next available judge was substituted for a missing one. For instance, if the top-ranked judge from the prior trial did not make a forecast in the current trial, we selected the judge ranked second instead.

Results

Figures 1 and 2 plot the expected performance of the judgment strategies relative to the average judge (*y*-axis) against varying levels of *k* (*x*-axis) for the experimental and economic datasets, respectively.

Experimental data. We first compared averaging and best-member judgment strategies. Across the 40 datasets, averaging outperformed the average judge by 37% (*M* = .63, *SD* = .13), whereas the best-member strategy outperformed the average judge by 12% (*M* = .88, *SD* = .24) when selected on recent performance and by 25% (*M* = .75, *SD* = .24) when selected on cumulative performance. Accordingly, averaging performed significantly better than the recent best member overall, 95% CI [.18, .33], and in 35 of the 40 datasets.[4] Averaging also performed significantly better than the cumulative best member, 95% CI [.04, .20], a pattern true in 27 of the 40 datasets. The superior performance by cumulative over recent best members was also significant, 95% CI [.05, .21].

Next we contrasted the performance of best-member and averaging strategies to the average performance of the small-crowd strategies (i.e., the mean performance for $k > 1$). When selected on recent performance, the average small crowd (*M* = .66, *SD* = .16) was significantly more accurate than best members, 95% CI [.17, .27], and as accurate as averaging, 95% CI [–.01, .08]. When selected on cumulative performance, the average small crowd (*M* = .58, *SD* = .14) was significantly more accurate than best members, 95% CI [.10, .22], and as accurate as averaging, 95% CI [–.08, .001]. In short, for $k > 1$, small crowds on average outperformed best

[4] Given the hierarchical structure of the data (datasets nested within samples), multilevel modeling was used for all inferential tests (random-intercepts only). Confidence intervals are for the difference in mean performance between strategies.

members and performed as well as averaging the entire crowd. On average, small crowds selected on cumulative performance outperformed those selected on recent performance, 95% CI [.03, .12].

As Figure 1 illustrates, there was considerable heterogeneity in small-crowd performance depending on *k* (SC-R: Wald $X^2(8) = 206.15$, $p < .001$; SC-C: Wald $X^2(8) = 46.04$, $p < .001$). So one natural question is how many opinions are enough? In this data, it depended on available history. With only recent performance, small crowds of five ($M = .66$, $SD = .17$) were necessary to perform as well as averaging, and no small crowd outperformed averaging. With cumulative performance, small crowds of two ($M = .64$, $SD = .19$) were as good as averaging, and small crowds of seven ($M = .56$, $SD = .15$) outperformed it, 95% CI [.00, .11].

Economic data. We first compared averaging and best-member judgment strategies. Across the six indicators, averaging outperformed the average judge by 21% ($M = .79$, $SD = .03$), whereas the best-member strategy outperformed the average judge by 15% ($M = .85$, $SD = .06$) when selected on recent performance and by 10% ($M = .90$, $SD = .08$) when selected on cumulative performance. Accordingly, averaging performed significantly better than the recent best member overall, 95% CI [.01, .11], and in five of the six indicators. Averaging also performed significantly better than the cumulative best member, 95% CI [.06, .16], a pattern true for all six indicators. The superior performance by recent over cumulative best members was not significant, 95% CI [–.00, .10].

Next we contrasted the performance of best-member and averaging strategies to the average performance of the small-crowd strategies (i.e., the mean performance for $k > 1$). When selected on recent performance, the average small crowd ($M = .74$, $SD = .04$) was significantly more accurate than best members, 95% CI [.08, .13], and averaging, 95% CI [.02, .07]. When

selected on cumulative performance, the average small crowd (M = .76, SD = .03) was significantly more accurate than best members, 95% CI [.08, .19], and as accurate as averaging, 95% CI [–.00, .05]. In short, for $k > 1$, small crowds on average outperformed best members and performed as well as averaging the entire crowd. The superior performance of small crowds selected on recent performance over those selected on cumulative performance was not significant, 95% CI [–.01, .05].

As Figure 2 illustrates, there was considerable heterogeneity in small-crowd performance depending on k (SC-R: Wald $X^2(8) = 60.26$, $p < .001$; SC-C: Wald $X^2(8) = 128.10$, $p < .001$). When selected only on recent performance, small crowds of three (M = .75, SD = .04) were enough to outperform averaging, 95% CI [.01, .07]. Based on cumulative performance, small crowds six (M = .75, SD = .03) were necessary to beat averaging, 95% CI [.01, .07].

Discussion

Study 1 examined the performance of small crowds in estimates made by experimental participants and professional forecasters. Small crowds outperformed best members regardless of the performance history available, with considerable effect size. On average, they also performed as well as averaging all judges, and in many cases managed to outperform it.

Having established the normative promise of using small crowds in this initial study, we turn in the remaining studies to peoples' intuitions about the effectiveness of the small-crowd strategy. In Study 2, we asked students to either rate or rank five strategies for using the forecasts of professional economists to make future predictions. We find that people prefer a small crowd to averaging the entire group and chasing best members.

Study 2

Method

Sixty-four students from two universities in the southeastern United States participated in exchange for a small payment. All were shown the following information:

> Every six months the *Wall Street Journal* surveys a panel of approximately 50 economists. The journal asks these economists to make forecasts of several economic statistics, including unemployment, the inflation rate, etc. At the end of the six months, 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 3-month Treasury Bill rate. This is the rate of interest the U.S. government pays for 3-month loans (the annualized rate as of March 2, 2007 was 5.03%). 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 Bill rate six months from today.

At this point, 40 participants were asked to rate the accuracy of the following five judgment strategies (1 = *not at all accurate* to 7 = *extremely accurate*), which were presented in one of two random orders: (1) the forecast of a randomly chosen economist; (2) the average forecast (the mean) of the entire panel of economists; (3) the forecast of the most accurate economist from the last period; (4) the forecast of the cumulatively most accurate economist over many periods; and (5) the average forecast (the mean) of the five most accurate economists from last period. The remaining 24 participants were asked to rank the same strategies based on their expected accuracy (1 = the most accurate strategy). Upon completion, participants were paid and dismissed.

Results and Discussion

Table 2 presents the average rating for each strategy. A repeated-measures analysis of variance indicated significant differences in perceived accuracy across judgment strategies, $F(4, 39) = 42.75$, $p < .001$, $\eta^2 = .45$. On average, the small-crowd strategy was rated alongside the cumulative expert as the most accurate strategies, followed by the recent expert, averaging the entire panel, and using a random economist. Table 3 summarizes the ranking of judgment strategies. An analysis of variance for ranked data (Winer, Brown, & Michels, 1991) indicated significant differences in perceived accuracy across strategies, $\chi^2(4) = 37.03$, $p < .001$. Averaging a small crowd had the lowest mean rank (i.e., highest perceived accuracy) and was ranked first by 10 of 24 participants, followed closely by the cumulative expert.

Given the choice, people in this study preferred to average the opinions of a small group of high performers over chasing best members or averaging the entire group. This was mirrored in the ratings of perceived accuracy for each strategy. Study 2 thus provides preliminary evidence that, in addition to their promise as a normative strategy for combining judgments, small crowds have intuitive appeal to decision makers. We made the availability of this strategy deliberately salient in these studies, so it remains to be seen if unprompted decision makers also take advantage of small crowds in a tasks with real stakes. This is addressed in Study 3, in which participants made a series of predictions based on actual forecasts from a panel of economists and were rewarded for accuracy.

Study 3

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* semi-annual surveys. Participants were rewarded for the accuracy of their forecasts, and our primary

dependent measure was whether they chose to average the economists' forecasts, average the forecasts of a small crowd, or rely on a best-member strategy.

Method

Sample. Eighty students from a university in the southeastern United States (M_{age} = 20.6 years, 42 female) participated in exchange for a fixed fee of $5 and a bonus based on performance (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. For each of 24 total periods, participants in the *low-choice* condition could choose either the forecast of one economist or the average forecast of all 11 economists (i.e., they could only choose best members or average the panel), whereas participants in the *high-choice* condition could choose the forecast of one economist, the average forecast of all 11 economists, or the average forecast of any subset of economists (i.e., a small-crowd strategy). Participants in the *low-feedback* condition were provided feedback about performance (their own and that of the economists) in the prior period only, whereas participants in the *high-feedback* condition were provided feedback about both last period's performance and cumulative performance (equally weighted over the periods).

Eleven economists were chosen from the *Wall Street Journal's* semi-annual survey of forecasters who were present for all 24 surveys conducted from January 1994 to July 2005. Each provided predictions for the interest rate on 3-month U.S. Treasury Bills (T-Bills) six months from the survey date. The economists and predictions were the same 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.)

Procedure. At the beginning of each period, participants saw a table with the names of the 11 economists listed in rows alongside their forecasts of the T-Bill rate for the most recent period, the actual T-Bill rate, their forecast errors (in absolute value), and their average forecast errors (in absolute value) over all prior periods for participants assigned to the high-feedback condition. The 12th row in the table was for the "Average Forecast," which was the arithmetic mean of the 11 economists' forecasts. The last row of the table summarized the participant's ongoing performance.

After reviewing this information, participants selected their economists for the next period. Those in the low-choice condition could select either one of the 11 economists or the Average Forecast. Participants in the high-choice condition could select one economist, the Average Forecast, or any subset of two or more economists (selecting all 11 economists was coded as choosing the Average Forecast). If they chose multiple economists, the evaluated forecast was the arithmetic mean of those selected. Note that participants were not making their own predictions; they were relying on the single or average prediction of the economists. Moreover, they did not see the economists' forecasts for the next period; their selection was based solely on feedback about performance.

Once finished with the 24 periods, participants answered questions about the expected performance of different forecasting strategies and their beliefs about bracketing. They then learned their bonuses, were paid, and dismissed.

Results

Strategy selection. We coded selecting one economist as a best-member strategy, 2 to 10 economists as a small-crowd strategy, and all 11 economists as averaging. Figure 3 plots the number of economists selected on the *x*-axis and the frequency of that choice on the *y*-axis. In

the low-choice condition, the typical preference for best members over averaging is starkly apparent: Participants chose best members 80.0% of the time and averaging 20.0% of the time. The typical person in this condition averaged in only two of the 24 periods, and there were only four participants who averaged more than half the time.

Participants in the high-choice condition chose best members 13.0% of the time, averaging 27.6% of the time, and small crowds (ranging from two to seven economists) 59.4% of the time. Not surprising, there was considerable experimentation in this condition, with 28 of the 40 participants trying all three strategies. But two strategies comprised more than half the choices: a small crowd of two economists (28.7%) and averaging the entire panel (27.6%).

Figure 3 clearly indicates that when given the option, people shifted their choices from best members to small 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 best members or averaging was modeled a dichotomous dependent variable, and the high-choice condition, in which the choice of best members, small crowds, or averaging was modeled as a trichotomous dependent variable. Feedback, a linear effect of period, and their interaction were included as predictors, and standard errors were adjusted to account for the clustering of observations within person. Feedback and period had neither collective nor individual effects on whether people chose to average in the low-choice condition, $\chi^2(3) = 4.13$, $p = .25$, which was not surprising given the low use of that strategy. Feedback and period did influence behavior in the high-choice condition, however, $\chi^2(6) = 11.49$, $p = .074$. Namely, participants receiving high feedback were more likely on average to choose best members over small crowds, $\chi^2(1) = 6.65$, $p = .010$. Feedback and period did not affect the choice between best

members and averaging, $\chi^2(3) = 2.50$, $p = .48$. These results are consistent with the patterns in Figure 3.

Performance. We include as a benchmark the performance achievable by all participants had they simply averaged the entire panel each period. The average forecast error of this strategy was .48 (*SD* = .46), which was superior to all but three economists on the panel over the 24 periods. A 2 (choice) x 2 (feedback) between-subjects ANOVA of performance revealed only a main effect of choice, $F(1, 76) = 5.06$, $p = .027$, $\eta_p^2 = .07$. On average, participants in the high-choice condition made lower forecast errors (*M* = .50, *SD* = .04) than those in the low-choice condition (*M* = .52, *SD* = .05). Neither the amount of feedback nor its interaction with choice affected performance. Only 16 of the 80 participants beat a policy of averaging every period.

We also examined period-level performance conditioned on the strategy selected. Figure 4 plots by choice condition the average forecast error on the *y*-axis against the number of economists selected on the *x*-axis (the patterns were unaffected by feedback, which we did not consider further). In the low-choice condition, averaging led to lower forecast errors (*M* = .44) relative to best members (*M* = .55), $F(1, 39) = 5.71$, $p = .022$. Strategy also affected performance in the high-choice condition, $F(6, 39) = 3.44$, $p = .008$. Individual comparisons (based on unadjusted *p*-values) indicated that best members were significantly outperformed by 3-person crowds ($t = 3.33$, $p = .002$), 4-person crowds ($t = 3.16$, $p = .003$), and averaging ($t = 2.47$, $p = .018$). No small-crowd selection significantly out- or under-performed averaging.

Discussion

The main and encouraging finding from this study is that when the option was made available, people shifted from using best members to small crowds and improved their

performance for it. We suspect this was not due to a belief in the wisdom of crowds, for averaging the entire panel was seldom used. Only in the first period, which lacked information about the relative expertise of the economists, did the majority of people average (66 of 80). Over the subsequent 23 periods, in the presence of feedback, the number of participants out of 80 who averaged in a round ranged from 4 to 24, with a median of 17—clearly a minority. In contrast, the number of participants out of 40 who selected small crowds after the first period ranged from 17 to 30, with a median of 25. These results are consistent with the positive evaluations of these judgment strategies in Study 2.

We were surprised that so few participants outperformed averaging. The results of Study 1 indicated that small crowds do very well compared to averaging, so at the very least we expected participants in the high-choice condition to compete. One potential explanation for this is that for this specific panel of economists, small crowds do not do as well as averaging. This was true of participants' modal strategy, which was choosing two economists. A policy of strictly selecting the two top-ranked economists from the prior period produced an average forecast error of .51 (versus .48 for strictly averaging) and outperformed averaging in only 12 of the 23 trials. However, had participants consistently selected the top three economists from the prior period, their average forecast error would have been .47 and they would have outperformed averaging in 16 of the 23 periods. And the policy that produced the smallest average forecast error (.46) was a small crowd of six economists—a strategy chosen only 14 times. In general, participants chose too few economists.

It was not readily apparent why participants switched their preferences from best members to small crowds in the high-choice condition. It may be that people appreciated the benefits of averaging but were averse to including the least qualified judges. Thus small crowds

allowed them to both chase expertise and capitalize on the error-correction process of averaging. We doubt this explanation for two reasons. Prior evidence has established that people in general do not understand the benefits of averaging (Larrick & Soll, 2006). Moreover, if this explanation were true, we would expect to see small crowds of greater size. That is, if the appeal of small crowds is eliminating the worse judges from the group average, then small crowds of five to eight economists, for example, would appear more often than small crowds of two to three economists, as observed here. Alternatively, participants may have had no intuitions about small crowds per se, but simply experimented with different strategies and responded to the feedback they received about performance. This cannot be ruled out by the data in this study. People clearly responded to feedback and experimented, but 24 of the 40 participants in the high-choice condition settled on one strategy (e.g., best member, two economists, three economists, etc.) for at least half their judgments.

In Study 4 we test a third explanation. The fact that so many participants relied on two economists suggests that people prefer small crowds, not because they appreciate the value of averaging, but because they are uncertain about predicting best members. They therefore "hedge their bets" by selecting more than one.

Study 4

Method

Sample. Adults from a national panel (N = 211) completed an online survey in exchange for a small payment. Eight respondents provided incorrect answers to two or more questions in the training phase of the exercise and were excluded from all analyses. Two additional respondents were deemed inattentive and also excluded, leaving a final sample of 201.

Procedure. The survey 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 which must regularly predict attendance at upcoming events (full details are available from the corresponding author). Participants who incorrectly answered two or more of the questions were allowed to complete the study but were excluded from the analysis. In the evaluation phase, participants were presented with the following scenario:

> Now that we have reviewed miss size as a measure of accuracy, we would like to ask you a few questions about a recent film committee. The names of the five committee members are Amy, Brad, Carrie, Doug, and Eric. The table below shows how accurate each member was on average over the 30 most recent films:

	Amy	Brad	Carrie	Doug	Eric
Average Size of Miss	20	25	30	35	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 survey 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, $\chi^2(40) = 43.48$, $p = .33$ (maximum-likelihood estimate with oblique rotation). Questions A11, A12, and A13 were averaged and reverse-coded 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*; $\alpha = .66$); 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 and reverse-coded for an indicator of the perceived unpredictability of individual performance (*performance is unpredictable*; $\alpha = .87$); and S01, S02, and S03 were averaged and reverse-coded for an indicator of the perceived simplicity of averaging (*averaging is easier to use*; $\alpha = .82$). Including question A22 in *averaging reduces error* reduced the reliability of that measure to .53, so it was dropped from consideration. Note, higher scores on these measures indicate beliefs that favor averaging.

Strategy. Across the two orders (*choose–rate* vs. *rate–choose*), 44 participants chose to rely on the forecast of only one committee member (22.9%), 82 chose to average the forecasts of

two members (40.8%), 59 chose to average the forecasts of three members (29.4%), 5 chose to average the forecasts of four members (2.5%), and 11 chose to average the forecasts of all five members (5.5%). These choices were reduced to one of the three strategies: best members (22.9%), small crowds (72.6%), and averaging (5.5%). The use of small crowds was marginally higher in the *rate–choose* condition ($p = .094$, Fisher's exact test).

To understand what drove these choices, we conducted a multinomial logistic regression of strategy on the five measured variables, controlling for task order. The pseudo-R^2 of this model was .086, $\chi^2(12) = 25.06$, $p = .015$. Global tests of the joint significance of each predictor identified *averaging is more accurate*, $\chi^2(2) = 5.17$, $p = .075$, and *performance is unpredictable*, $\chi^2(2) = 5.19$, $p = .075$, as relevant to the choice of strategy (p-values for the remaining measured variables exceeded .35). All else equal, beliefs about the accuracy of averaging predicted when participants preferred to average rather than rely on best members ($B = 0.68$, $SE = 0.30$, $z = 2.27$, $p = .023$). The preference for small crowds over best members, in contrast, was driven primarily by beliefs about the predictability of individual performance ($B = 0.31$, $SE = 0.14$, $z = 2.23$, $p = .026$). These effects are illustrated in Figure 5.

Discussion

The results of Study 4 indicate that the choice to use small crowds rather than best members primarily reflects doubts by people about the predictability of future performance. The less people considered the past performance of these committee members a reliable guide to their future performance, the more likely they were to "hedge" by selecting more than one member to assist with the forecast. This is starkly illustrated in the bottom panel of Figure 5. The preference for a small crowd, moreover, was not meaningfully affected by beliefs about the benefits of averaging. As the top panel of the figure indicates, participants who understood that

combining opinions can improve accuracy simply averaged the forecasts of the entire panel instead of relying on Amy alone or a small crowd.

General Discussion

This research represents an initial attempt to establish the virtues of small crowds. We have identified two. Study 1 established in experimental and real-world datasets that small crowds are unequivocally wiser than best members and as wise if often wiser than the entire crowd. Small crowds elevate the mean expertise of a group compared to averaging all *and* capitalize on the error-correction process which best members cannot. Additional work is clearly needed in this area, both with empirical and simulated data. The latter is particularly well-suited to identifying the judgment environments in which small crowd are expected to succeed. Variables to be investigated include the dispersion in expertise, independence of the judges, and the diagnosticity of past performance. It will also be important in time to assess small-crowd performance on categorical judgments in addition to the judgments of quantity investigated herein.

The second virtue of small crowds is their attractiveness to decision makers. When faced with a choice between relying on an expert or on the average judgment of a crowd, people overwhelmingly prefer the former. In all but the most stable and predictable of environments, this will turn out suboptimal. But the choice between best members and averaging is a false one, for people can seek a second, third, even four or more opinions. Moreover, small crowds are forgiving of a person's inability to identify best members; all it requires is that he or she can identify the better members of a group. Studies 2 and 3 suggest that this alternative to best members and averaging will be exploited when salient, though perhaps not to its fullest potential.

We note two limitations of this series of studies and opportunities for future research. First, we have focused exclusively on the expected and actual performance of small crowds, but two other factors are very relevant to its evaluation and effectiveness. First, it is important to consider the variability in performance associated with each strategy. Intuitively, a best member strategy will be highly variable. At times the expert may reliably repeat in subsequent periods, but in other cases, achievement in one period will be followed by failure in another. In contrast, averaging the crowd will be least variable, often outperforming most judges but at worse performing no worse than the average judge. The variability of small crowds is expected to lie between these extremes, but it remains an empirical question.

Second, it is important to consider the costs associated with a small-crowd strategy (Payne et al., 1988). Each strategy varies in the number of "cognitive operations" needed to execute it. For example, the least costly strategy is to pick a judge at random each period. Averaging requires collecting members' opinions each period and calculating their mean, so it is a more demanding strategy. Small crowds require collecting and evaluating information about judge expertise, assembling opinions, and calculating a mean, so it is more costly still. Understanding the complexity of using these strategies, alongside their expected performance and variability, is an important consideration in determining the conditions under which each is recommended.

Conclusion

In general, research on benefiting from the wisdom of others has proceeded in two directions. Groups and teams scholars focus on how groups can better identify and use the expertise of their best members. Judgment and decision-making researchers, in contrast, focus on extracting the collective wisdom of crowds through aggregation techniques like averaging.

Although the normative superiority of the latter approach has been well demonstrated, the behavioral evidence overwhelmingly demonstrates people's preference for the former in practice. We have provided in this research initial evidence for the normative and behavioral appeal of small crowds. Not only does averaging the opinions of a small crowd of experts perform well compared to proven strategies like averaging, it also satisfies people's robust desire to chase expertise. As such, the wisdom of small crowds as a judgment strategy is a virtuous balance of prescriptive and descriptive considerations.

References

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgments. In H. S. Guetzkow (Ed.), *Groups, leadership and men: Research in human relations* (pp. 177-190). Pittsburgh, PA: Carnegie Press.

Bates, J. M., & Granger, C. W. J. (1969). Combination of forecasts. *Operational Research Quarterly, 20,* 451-468. doi: 10.2307/3008764

Baumann, M. R., & Bonner, B. L. (2004). The effects of variability and expectations on utilization of member expertise and group performance. *Organizational Behavior and Human Decision Processes, 93,* 89-101.

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.

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.

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

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.

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.

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.

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.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84*, 158-172.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117-140.

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

Galton, F. (1907). Vox populi. *Nature, 75*, 450-451. doi: 10.1038/075450a0

Gerena-Morales, R., & Hilsenrath, J. E. (2006, January 3). Pricey jeans give Hanmi CEO leg up to top rank of forecasters, *The Wall Street Journal*.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take-the-best heuristic. In G. Gigerenzer, P. M. Todd & ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 75-95). New York: Oxford University Press.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*, 149-167.

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.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (Vol. 2, pp. 129-157). London: JAI Press, Inc.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review, 112*, 494-508.

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

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

Hogarth, R. M. (1978). Note on aggregating opinions. *Organizational Behavior and Human Performance, 21*, 40-46.

Kaplan, M. F., & Miller, C. E. (1987). Group decision making and normative versus informational influence: Effects of type of issue and assigned decision rule. *Journal of Personality and Social Psychology, 53*, 306-313.

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.

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

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*, 177-189.

Libby, R., & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Decision Processes, 21,* 121-129. doi: 10.1016/0030-5073(78)90044-2

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

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.

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.

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

Maines, L. A. (1996). An experimental examination of subjective forecast combination. *International Journal of Forecasting, 12*, 223-233. doi: 10.1016/0169-2070(95)00623-0

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.

Miner, F. C. (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.

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.

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). NY: Psychology 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 Classics.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes, 84*, 288-307. doi: 10.1006/obhd.2000.2926

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: Academic Press.

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: Little, Brown.

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

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

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

Table 1

Economic Indicators Collected for Study 1

Statistic	Years	Periods	Forecasts	Median forecasts	
				Per period	Per Judge
Nominal GNP/GDP	1969-2009	161	6,117	35	13
GNP/GDP price index	1969-2009	161	6,083	35	14
Nominal corporate profits	1969-2005	147	4,672	28	12
Real personal consumption	1982-2009	110	3,527	33	10
Real residential investment	1982-2009	110	3,472	32	10
Real change in inventories	1982-2009	110	3,455	33	10

Table 2

Ratings of Judgment Strategies in Study 2

Strategy	*M*	*SD*	Difference in means			
			1	2	3	4
1. Random economist	2.35	1.12	–			
2. Average of entire panel	4.28	1.80	1.93**	–		
3. Recent expert	4.40	1.34	2.05**	0.12	–	
4. Cumulative expert	5.22	1.29	2.87**	0.94	0.82**	–
5. Average of small crowd	5.22	1.25	2.87**	0.94*	0.82	0.00

Note. $N = 40$. *$p < .05$ **$p < .01$ (Bonferroni-adjusted, $\alpha_{FW} = .05$)

Table 3

Ranking of Judgment Strategies in Study 2

Strategy	*M*	*SD*	Frequency ranked				
			1st	2nd	3rd	4th	5th
Random economist	4.50	1.25	2	1	0	1	20
Average of entire panel	3.00	1.14	2	8	3	10	1
Recent expert	3.21	0.98	1	5	7	10	1
Cumulative expert	2.38	1.31	9	3	8	2	2
Average of small crowd	1.92	0.93	10	7	6	1	0

Note. $N = 24$.

Table 4

Questions Associated with the Choice of Committee Members in Study 4

No.	Question
A11.	Averaging the forecasts of multiple committee members will be less accurate than choosing Amy to provide the forecast.
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.
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.*
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.
R22.	It is clear who the best forecaster will be in the future.
R23.	The past performance of each forecaster is a reliable guide to their future performance.
S01.	Choosing one committee member is a simpler strategy to use than averaging the forecasts of multiple committee members.
S02.	Averaging the forecasts of multiple committee members is more complicated than relying on the forecast of one member alone.
S03.	It is easier to go with one committee member than to average the forecasts of multiple committee members.

Note. Answered on 7-point Likert scale anchored by *strongly disagree* and *strongly agree*.

*Reverse-coded.

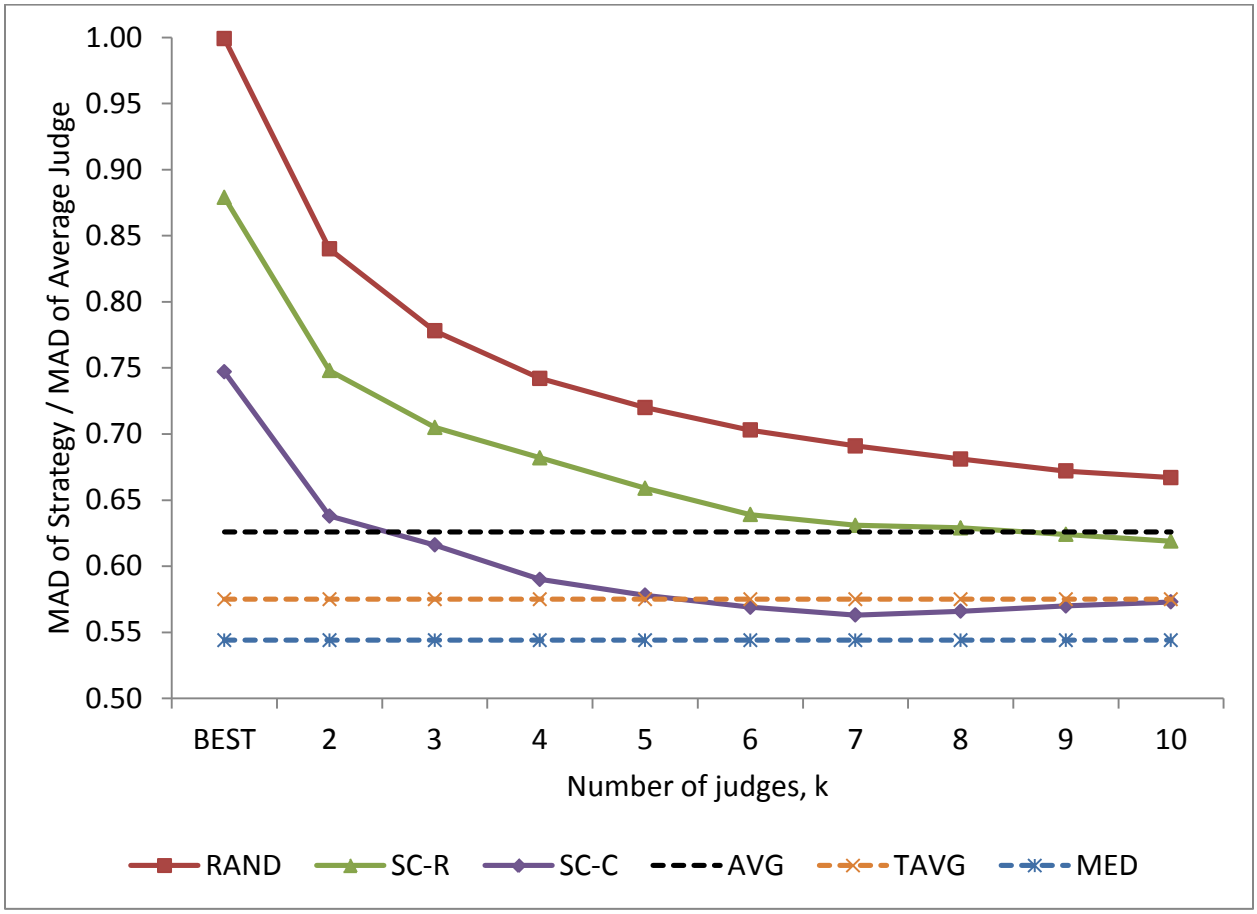

Figure 1. Performance of judgment strategies for the experimental data in Study 1. The *y*-axis is error relative to the average judge, so lower values indicate better performance. The performance of best members is illustrated at $k = 1$. Curves are shown for selecting judges at random (RAND), based on last period's performance only (SC-R), and based on cumulative performance (SC-C). The performance of the entire panel's mean (AVG), trimmed mean (TAVG), or median (MED) forecasts are shown as dashed lines.

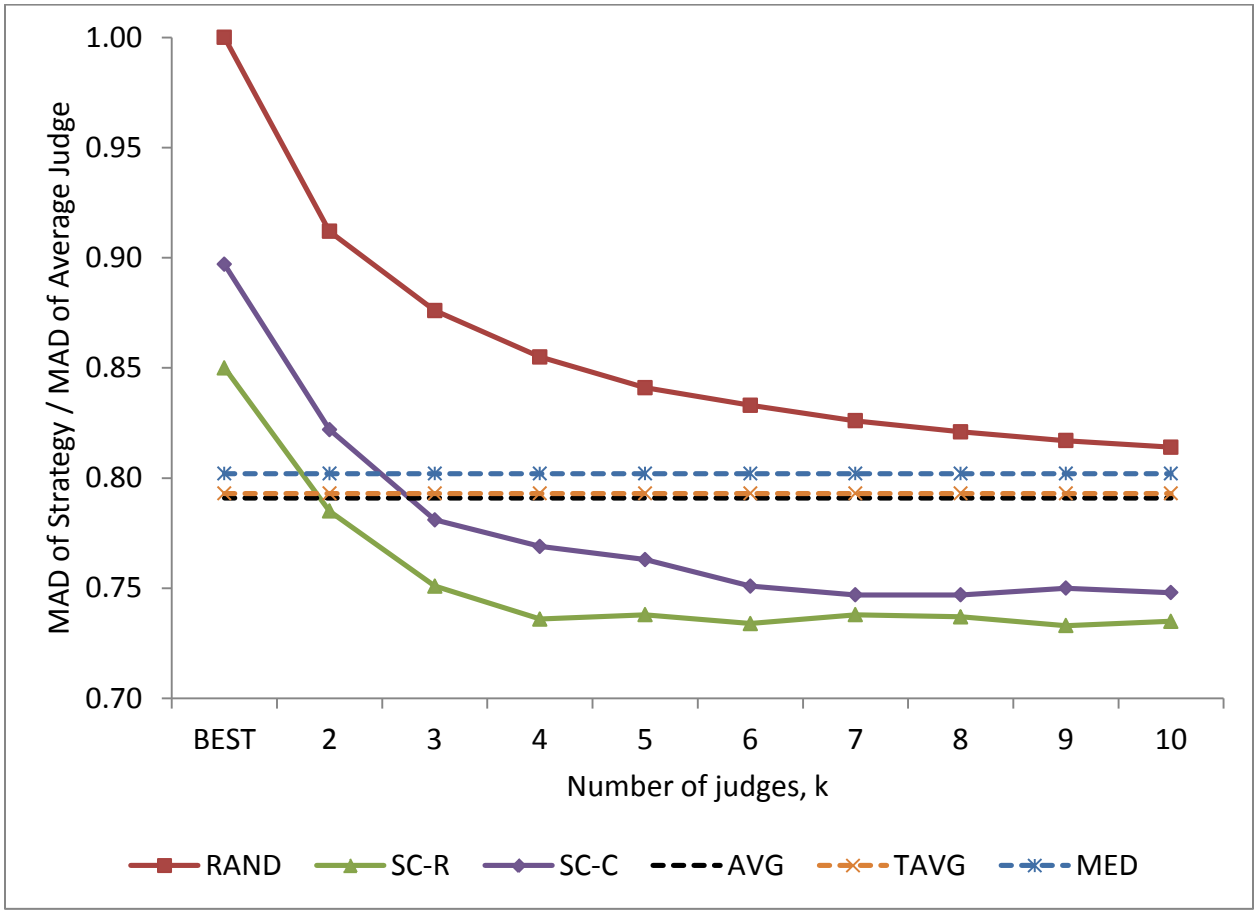

Figure 2. Performance of judgment strategies for the economic data in Study 1. The *y*-axis is error relative to the average judge, so lower values indicate better performance. The performance of best members is illustrated at $k = 1$. Curves are shown for selecting judges at random (RAND), based on last period's performance only (SC-R), and based on cumulative performance (SC-C). The performance of the entire panel's mean (AVG), trimmed mean (TAVG), or median (MED) forecasts are shown as dashed lines.

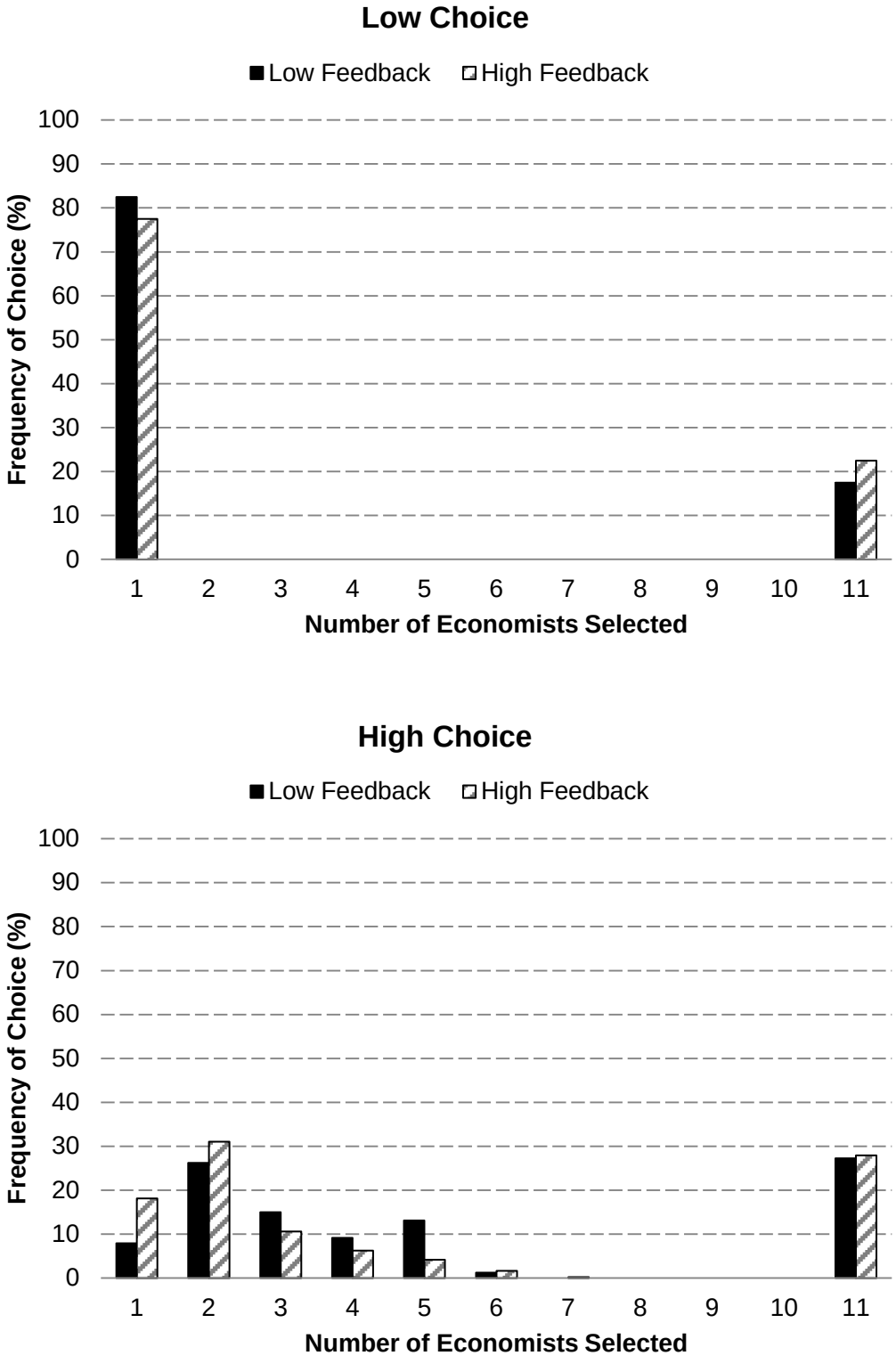

Figure 3. Frequency of judgment strategies across periods in Study 3 by condition.

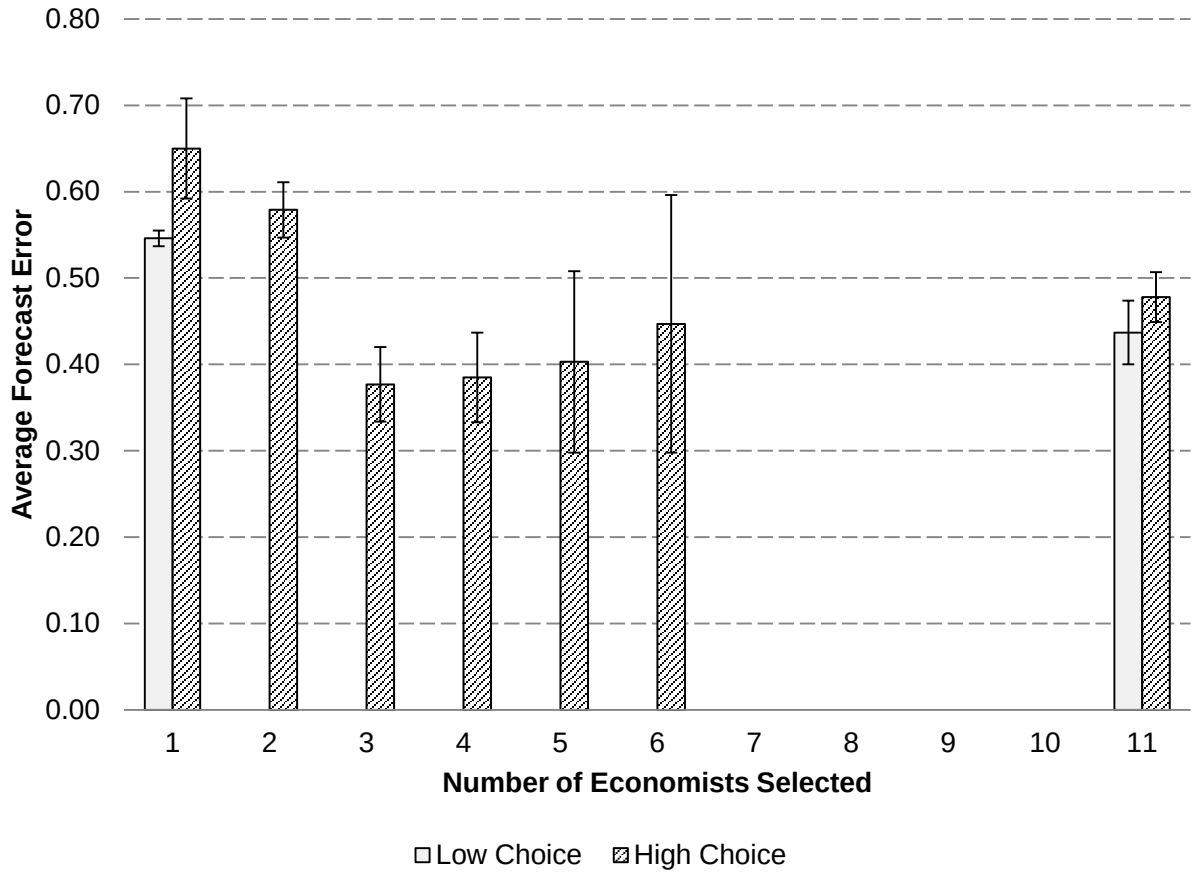

Figure 4. Period-level performance conditioned on strategy in Study 3. Means with standard errors are reported. Lower values indicate better performance.

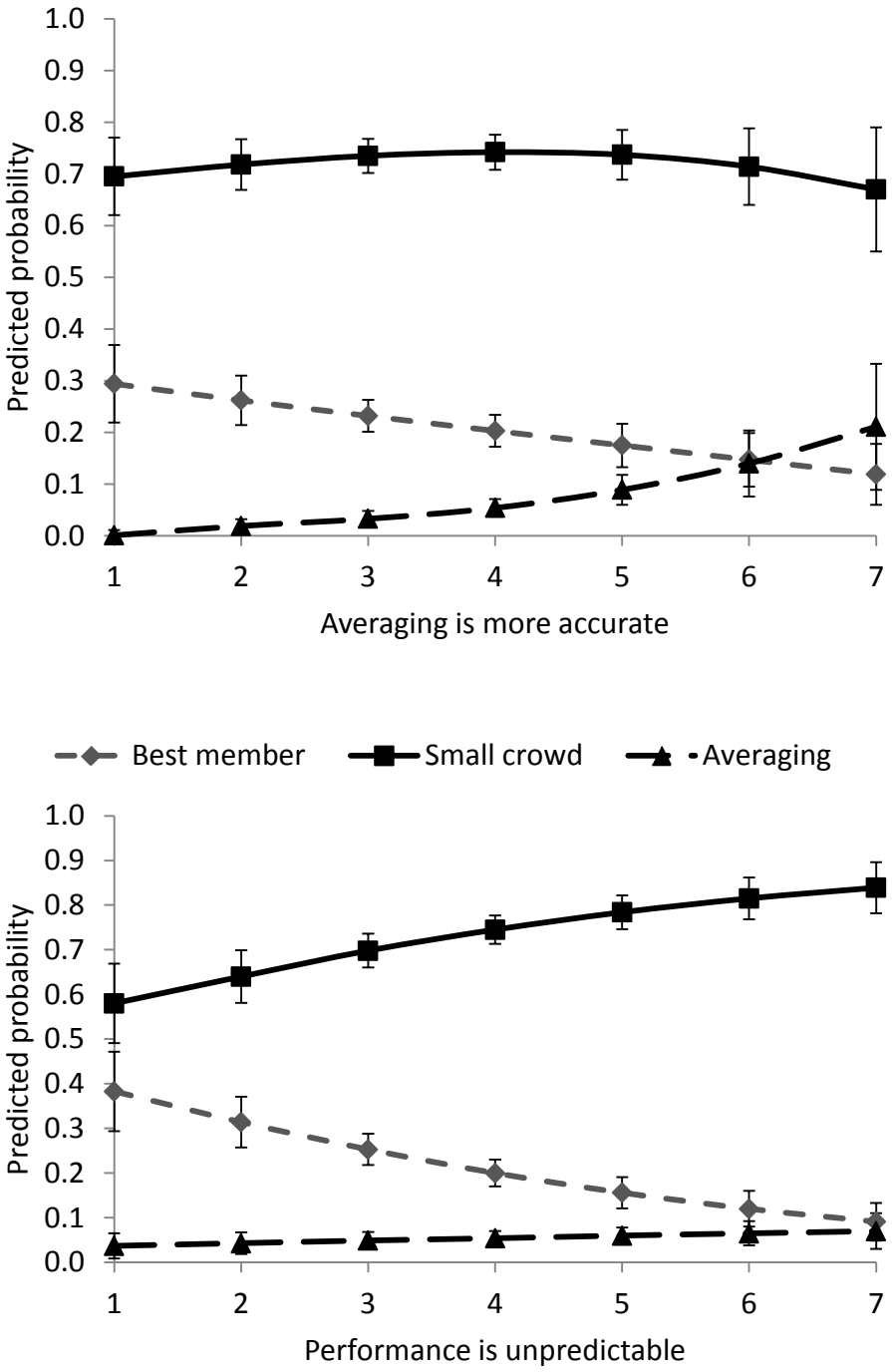

Figure 5. Predictors of the use of small crowds and averaging in Study 4. The likelihood of averaging increased with perceptions of its accuracy (top panel). The likelihood of small crowds increased with perceptions that individual performance is unpredictable (bottom panel).