Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision

Albert E. Mannes
Fuqua School of Business, Duke University, Durham, North Carolina 27708, aem15@duke.edu

Recent research has advanced our understanding of how people use the advice of others to update their beliefs. Because groups and teams play a significant role in organizations and collectively are wiser than their individual members, it is important to understand their influence on belief revision as well. I report the results of four studies examining intuitions about group wisdom and the informational influence of groups. In their overt assessments, experimental participants rated larger groups as more accurate than smaller groups and discriminated more between them when group size was salient. When provided advice, participants relied more on groups than individuals to update their beliefs, but were only modestly sensitive to group size. Most were suboptimal in the use of that advice, overweighting their initial beliefs and underweighting the more valid judgment of the group. Thus although acknowledged in principle, the wisdom of crowds is only shallowly manifest in observed behavior.

Key words: organizational studies; decision making; groups; combining judgments; influence; advice-taking

History: Received July 24, 2008; accepted March 26, 2009, by George Wu, decision analysis. Published online in Articles in Advance June 26, 2009.

1. Introduction
In this increasingly interconnected world, how individuals use the opinions of others remains an important topic for decision-making researchers. In regard to the use of numerical judgments (e.g., a sales forecast, probability estimate, etc.), recent research describing intuitive combining strategies generally follows one of two paradigms. First is the panel-of-experts paradigm, in which individuals form a judgment on the basis of two or more expert opinions (Budescu 2006, Budescu and Rantilla 2000, Budescu et al. 2003, Fischer and Harvey 1999, Maines 1996, Yaniv 1997). The consensus from this research is that people tend to average these opinions but err in their assignment of weights. Second is the advice-taking paradigm, in which individuals update their prior beliefs with the opinion of one or more advisors (Bonaccio and Dalal 2006, Harvey and Fischer 1997, Harvey and Harries 2004, Harvey et al. 2000, Koechler and Beauregard 2006, Lim and O’Connor 1995, Soll and Larrick 2009, Yaniv 2004, Yaniv and Kleinberger 2000, Yaniv and Milyavsky 2007). This work has reached a similar conclusion: People use advice to improve their beliefs but improperly weight judgments of the self and advisor(s). In particular, people tend to favor their own, initial beliefs and underweight advice.

One dimension neglected by both these traditions is the effect of collective group judgments on individual beliefs. That is, what is the differential impact on our beliefs of advice provided by a group rather than an individual? This is important for multiple reasons. Teams are prevalent in modern organizations. In a recent survey, 48% reported using teams in some form, the average size of which was 11 people, and they were more likely in firms with multiple departments and revenues exceeding $100 million (Devine et al. 1999). Moreover, team and group judgments are frequently wiser than those of individuals (Einhorn et al. 1977, Gigone and Hastie 1997, Hastie 1986, Hill 1982), the case for which was recently popularized by Surowiecki (2004) in The Wisdom of Crowds. Thus, a normative argument can be made that group and team judgments should be more influential in general than individual judgments. Finally, research into the study of groups as producers of information (e.g., Hinsz et al. 1997) has not been accompanied by an understanding of the consumption of that information by individual decision makers. Although researchers using both the panel-of-experts and advice-taking paradigms have used multiple experts or advisors in their studies (e.g., Budescu 2006, Yaniv and Milyavsky 2007), they are presented as distinct social entities (cf. Wilder 1977) and offer individual opinions, not collective judgments. Thus, their research does not address this question directly.

In contrast, some of the more memorable experiments of social psychology in the last century have revealed the profound influence groups have on individual beliefs (e.g., Asch 1951, Sherif 1935). It is important to distinguish in this research, however, between
the normative and informational social influence of groups (Deutsch and Gerard 1955, Kelman 1958). The former refers to temporary compliance with a group judgment to satisfy the expectations of others, but is not accompanied by a permanent internalization of the group’s belief. A classic example is the willingness of people to go along with patently incorrect judgments by a unanimous majority about the length of a line (Asch 1951). The influence of present concern, however, is a group’s informational influence on individual beliefs, which refers to a permanent internalization of the group’s judgment (or a part thereof) to hold veridical beliefs about reality (Festinger 1954). Sherif’s (1935) studies of group-norm formation in judgments about the illusory movement of light induced by the autokinetic effect are a better example of informational influence. Nonetheless, as in the studies of normative influence, the stimuli and group judgments in this work were purposely designed to be deceptive and maladaptive (Allen 1965). Evidence for the informational influence of groups in tasks more representative of those encountered in natural social environments is lacking (Brunswik 1955, Tajfel 1969).

The purpose of the current research is to address this gap in our understanding of group informational influence. After reviewing relevant theory, I report the results of four experiments that assessed the influence of group judgments on individual beliefs. For this purpose I chose to use an advice-taking paradigm, which focuses on belief revision or updating instead of belief formation. In this paradigm, participants make a series of initial judgments and subsequently revise them aided by the additional judgment of an influence source.2 In past advice-taking research this influence source has been one or more distinct advisors, but the current research uses groups of various sizes. The participants’ final judgments are modeled as an additive linear combination of their initial beliefs and the group judgments. By examining the weights the participants place on each cue, I assess the relative influence of the group in participants’ judgment policies (cf. Hoffman 1960).

2 Theory and Hypotheses

2.1. Informational Influence of Groups

Although plagued by the confounding of normative and informational influence in the same studies (Campbell and Fairey 1989), research on social influence does predict that groups will be more influential than sole individuals on our beliefs. Formal models have consistently demonstrated that conformity to a source judgment increases as a monotonic function of source size (Bond 2005, Campbell and Fairey 1989, Latané 1981, Latané and Wolf 1981, Tanford and Penrod 1984). Thus, two people are more influential than one, three more than two, and so forth, although the exact functional form of the relationship varies by model and study.

In the case of informational influence, a possible explanation for this result is that people believe group judgments are more accurate than individual judgments—that is, they appreciate the wisdom of crowds. People are sensitive to asymmetries in expertise and place more weight on the perceived expert in their combining strategies (Birnbaum and Stegner 1979, Budescu et al. 2003, Harvey and Fischer 1997, Lim and O’Connor 1995, Maines 1996, Sniezek and Van Swol 2001, Yaniv and Kleinberger 2000). Therefore, as groups grow wiser with increases in size, albeit at a diminishing rate (Einhorn et al. 1977, Hogarth 1978, Libby and Blashfield 1978), people are expected to place more weight on their judgments. Direct evidence for this prediction comes from a recent study of attitude formation in which experimental participants high in accuracy motivation were more influenced by a fictitious consensus opinion of 1,000 respondents than by the identical opinion of 10 respondents and rated the former as more reliable than the latter (Darke et al. 1998).

HYPOTHESIS 1 (H1). The relative influence of group judgments on individuals’ final beliefs will increase monotonically with group size.

Yet there is reason to predict that people will make suboptimal use of group judgments. In an early experiment, Jenness (1932) had students guess the number of beans in a bottle, meet in 3-person teams to submit a consensus group estimate, and then provide updated individual guesses. The average absolute individual errors were 305 and 122 beans before and after team discussion, respectively. Clearly, meeting as a group improved individual judgment. However, the average absolute error of the team judgments was 91 beans! Seemingly, the students placed too much weight on their initial beliefs and too little on the team judgment in their final estimate. This is common in the advice-taking literature, where average weights on others’ opinions are in the 0.20–0.40 range and at the item level are frequently zero (Goldberg 1954, Harvey and Fischer 1997, Harvey and Harries 2004, Koehler and Beauregard 2006, Lim and O’Connor 1995, Soll and Larrick 2009, Yaniv 2004, Yaniv and Kleinberger 2000).

This bias in favor of one’s own beliefs does not arise from simple arithmetic errors during opinion aggregation, though these are surely present. In a study by Lim and O’Connor (1995), experimental participants could specify the weights for the linear combination

2 Source is used as a more general term to include both individuals and groups.
of judgments without having to make the actual calculation. If they knew the appropriate weights to assign, this decision-support system should have improved the accuracy of the composite judgments. Surprisingly, they found the opposite: participants specified weights that led to less accurate composites vis-à-vis a control group who combined the judgments intuitively. Instead, the bias may be explained by a combination of factors: a general conservatism in belief updating (Edwards 1982), anchoring or order effects (Hogarth and Einhorn 1992, Tversky and Kahneman 1974), access only to reasons for one’s own judgments (Yaniv and Kleinberger 2000), and undue attention paid to a focal hypothesis (in this case, one’s initial belief) (Schkade and Kahneman 1998, Wilson et al. 2000).

Even if people appreciate in principle the wisdom of crowds, this bias is not expected to disappear with group judgments. Because the relationship between the size of a group and the validity of its average judgment is a function of both members’ expertise and the correlations among their judgments (Hogarth 1978), applying the law of large numbers to any particular situation is difficult even for sophisticated judges. And in many cases people act in accordance with the “law of small numbers” instead (Tversky and Kahneman 1971). Accordingly, they are expected to be insufficiently sensitive to the size of the group in their judgment policies.

Hypothesis 2 (H2). The relative influence of group judgments will be increasingly suboptimal as group size increases.

To argue that people will make suboptimal use of a group’s informational benefits requires a normative standard. In terms of the linear model of judgment proposed in the introduction, this is often the policy specified by regressing the true values of the criterion on the predictor judgments. This yields cue weights that maximize the correlation between the linear composite and the criterion. However, a policy that approximates the performance of this yet is an arguably fairer, ex ante prescriptive standard is unit weighting (Armstrong 2001, Clemen 1989, Dawes and Corrigan 1974, Einhorn and Hogarth 1975). The arithmetic mean, for example, is a unit-weighted composite. Moreover, when one of the judgments is a group’s mean estimate, unit weighting is equivalent to a weighted average. That is, when combining a prior belief with the mean judgment of an N-person group, unit weighting implies a weight of $1/(N+1)$ on the prior belief and a weight of $N/(N+1)$ on the group’s judgment. For example, imagine a sales executive updating her own belief about future sales with the mean forecast of a 4-person analyst team. In total there are five opinions among them (that is, the sales executive’s forecast and the four forecasts summarized by the team’s mean). Unit weighting means that each opinion contributes 20% to the final judgment. Thus, in this case, the executive’s intuitively revised belief can be compared to a weighted-average composite that places 80% of the weight on the team’s mean judgment and 20% on the executive’s prior belief. For the remainder of this paper, this will be referred to as an ego-neutral judgment policy, in contrast to an egocentric judgment policy that overweights judgments of the self (Yaniv and Kleinberger 2000). To restate H2 in light of this standard, individuals’ intuitive judgment policies are expected to feature weights on the group that increasingly fall short of those implied by an ego-neutral judgment policy.

2.2. Benefiting from Group Judgments

Research on the influence of groups by and large has focused on its detrimental impact on individual judgments. In contrast, a contribution of this research is to show the benefits of yielding to a group’s informational influence. If people recognize the wisdom of crowds and place more weight on their judgments, the validity of their revised beliefs will increase. Part of these gains will be due simply to averaging one judgment with a nonredundant second judgment, but part will reflect the superior expertise of the groups. Thus, these improvements are predicted to be greater for larger, wiser groups.

Hypothesis 3 (H3). The influence of group judgments will improve individual beliefs monotonically with group size.

As discussed, intuitive judgment policies are expected to be egocentric compared with ego-neutral judgment policies. Whether this is consequential depends on the validity of the additional judgment. Individuals who receive advice from a sole individual typically combine the judgments with approximately 0.70/0.30 weights on the self and advisor, respectively (Harvey and Fischer 1997, Koehler and Beauregard 2006, Lim and O’Connor 1995, Yaniv 2004). With advisors of similar expertise, this deviation from the 0.50/0.50 weights of an ego-neutral policy will have little effect. This is because the greatest gain in validity is from one judgment to two (Einhorn et al. 1977, Hogarth 1978, Libby and Blashfield 1978), which will be robust to deviations from unit weighting (Dawes and Corrigan 1974). When advised by groups, however, the marginal contribution one’s own belief

---

3 The terms ego-neutral and egocentric in this sense are strictly descriptive and do not imply that motivational factors related to self-enhancement or ego-maintenance are necessarily affecting judgment policies. Also note that the weights implied by an ego-neutral judgment policy may be greater or less than optimal weights derived through regression.
can make to a composite judgment quickly declines. Overweighting it will lead to significant underperformance vis-à-vis an ego-neutral policy.

Hypothesis 4 (H4). The suboptimal influence of groups in belief revision will be more consequential with larger groups.

3. Overview of Studies
The advice-taking paradigm used in these experiments is well suited to the study of a group’s informational influence. Typical studies of group influence tend to confound normative and informational influence, making separate assessments of them problematic. The present studies incorporate a number of features designed to minimize normative pressures to conform and to maximize informational influence (Allen 1965, Campbell and Fairey 1989, Deutsch and Gerard 1955, Kaplan 1989, Tajfel 1969). Namely, experimental participants respond privately and anonymously to the group judgment; they are rewarded for accuracy, which motivates the search for truth; and the judgments are difficult but verifiable and ecological (i.e., they are not manufactured to be better or worse than those of the participants).

The results of three experiments are reported that examined the levels and consequences of group informational influence. The primary experimental factor was source size, manipulated between participants in two studies and within participants in one. Influence was measured using an idiographic, regression-based approach common in the decision-making literature (see Cooksey 1996, Hoffman 1960). Each participant made multiple quantity estimates with the help of two cues: his or her own initial belief and the judgment of an influence source. A multiple regression model for each participant summarized his or her judgment policy.4 Because the criterion and predictor judgments were similarly scaled, unstandardized regression coefficients captured the importance of each cue. Though not without shortcomings as a measure of importance (Bonaccio and Dalal 2006, Cooksey 1996), these coefficients were directly interpretable, preserving both the direction and relative strength of each cue’s influence. To test hypotheses about influence, coefficients on the influence source were compared across experimental factors. To assess how well participants profited from advice, the accuracy of their final estimates was compared with that of their initial estimates and the estimates formed by an ego-neutral judgment policy.

Leading off, though, is a preliminary study in which people were asked directly about their perceptions of the wisdom of crowds.

3.1. Study 1

3.1.1. Task and Procedure. Sixty-three students from a private U.S. university (40 female; $M_{\text{AGE}} = 20.4$ years) participated in the experiment in exchange for a participation fee of $5. Participants sequentially rated the validity of judgments made by 2- and 9-person groups on a 7-point Likert scale ($1 = \text{Extremely inaccurate}, 7 = \text{Extremely accurate}$). They were asked: “Imagine that you are making temperature predictions. You get an average estimate from a group of 2 [9] people (i.e., the sum of the estimates divided by 2 [9]). Over many predictions, how accurate do you believe the group’s estimate will be?” The order of ratings (i.e., 2-then-9 versus 9-then-2) was assigned randomly between participants.

3.1.2. Results and Discussion. Sample sizes were 35 and 28 in the 2-then-9 and 9-then-2 conditions, respectively. A repeated-measures analysis of variance indicated that the accuracy of 9-person group judgments ($M = 4.3, SD = 1.3$) was rated higher than that of 2-person group judgments ($M = 4.1, SD = 1.2$) ($F(1, 61) = 61.67$, $MSE = 0.740, p < 0.001; d = 0.95$). This within-participants difference was not significantly affected by the order in which the ratings were made. The same pattern also emerged between participants, albeit diminished, when comparing the ratings made first in each order condition: 9-person group judgments ($M = 4.0, SD = 1.3$) were rated as more accurate than 2-person group judgments ($M = 3.4, SD = 1.2$) ($F(1, 61) = 4.59$, $MSE = 1.49, p < 0.05; d = 0.54$).

In sum, when asked to compare 2- and 9-person groups directly, participants rated the latter as more accurate. The results suggest that people have the correct intuitions about the ordinal relationship between judgment validity and group size (Darke et al. 1998). Whether this was also true of people’s behavior was the subject of the remaining studies.

3.2. Study 2

3.2.1. Task and Procedure. Eighty students from a private U.S. university (48 female; $M_{\text{AGE}} = 19.9$ years) participated in the experiment in exchange for a participation fee of $4 and a bonus for accuracy ($M = 2.34, SD = 0.69$). The task was to estimate the average high temperature in January of 20 U.S. cities with the aid of an influence source. The size of the influence source—one, 2, 4, or 9 people—was manipulated between participants. These sizes were
chosen deliberately to facilitate use of an ego-neutral judgment policy (i.e., weights on the source of 0.50, 0.67, 0.80, and 0.90, respectively). The 20 cities were chosen by randomly generating five-digit numbers and matching those to U.S. postal zip codes. Fifteen separate students estimated the average high temperature in January (in degrees Fahrenheit) for these cities on a prior occasion in exchange for a small payment. Sources of size 1, 2, 4, and 9 people were randomly generated from this pool, and for each city the mean of source members’ temperature estimates was calculated. This was the advice presented to the participants (the individual estimates constituting the average were not presented). To ensure the sample estimates were representative of average across participants and to reduce bias introduced by any one sample, this process was repeated until there were 15 samples of 1-person sources, 15 samples of 2-person sources, and so forth; within condition participants received 1 of these 15 samples. As for the incentive to be accurate, each city was worth 25 cents, and for each degree Fahrenheit by which a participant’s estimate missed the truth, he or she lost 2 cents. Thus, participants could earn a $5.00 bonus for perfect estimates and were equally penalized for over- or underestimation. Though they could lose money at the item level, the minimum bonus over all 20 cities was truncated at zero. After passing a test to ensure comprehension of the bonus payment, participants proceeded to Part 1.

In Part 1 participants made their initial, unaided temperature estimates. The cities were presented sequentially and in random order to each person. Participants entered and confirmed their temperature estimates. The cities were presented sequentially and in random order to each person. Each trial included the city, the participant’s initial estimate, and the estimate of an influence source. These were described as “the estimate of another student” (1-person source) or “the average estimate of 2 [4, 9] students” surveyed on a prior occasion. Explanation and examples of “the average estimate” were provided prior to starting. Participants entered and confirmed their final estimate on the computer. They did not receive feedback on their performance during the task, nor could they return to prior estimates. Participants answered a few questions upon completion of the task, after which they were debriefed and paid.

3.2.2. Dependent Measures. Influence was compared for two judgment policies. For the observed judgment policy, final estimates were modeled as an additive linear combination of the initial, unaided estimates and the estimates of the influence source. The unstandardized regression coefficient on the source estimates measured the observed level of source influence (see Footnote 4). For the ego-neutral judgment policy, levels of source influence were set at 0.50, 0.67, 0.80, and 0.90 for the 1-, 2-, 4-, and 9-person sources, respectively. Performance was measured as the mean absolute deviation (MAD) between participants’ final (i.e., revised) estimates and the true temperatures across the 20 cities. MAD is a widely accepted measure of accuracy for symmetric loss functions (Armstrong 2001), with lower MADs indicating better performance. MADs were also calculated for each participant’s initial, source, and ego-neutral estimates.

3.2.3. Results.

Preliminary Analysis. Table 1 indicates that the expertise of participants (i.e., initial MADs) did not vary systematically across conditions ($F(3, 76) = 0.66$, $MSE = 28.22$, ns). Nevertheless, because initial MAD was correlated with both influence ($r = 0.34$) and final MAD ($r = 0.72$), it was included as a covariate in the analysis of these measures to increase power. As expected, the validity of the source increased with source size (i.e., source MADs declined with size) in accordance with the law of large numbers ($F(3, 76) = 17.17$, $MSE = 2.10$, $p < 0.001$; $R^2 = 0.40$).

Influence. The mean level of source influence was 0.54 ($SD = 0.20$), which varied by condition ($F(3, 75) = 9.19$, $MSE = 0.028$, $p < 0.001$; $R^2 = 0.35$). Table 1 indicates that source influence increased concavely with source size. An orthogonal trend analysis for unequal intervals (Keppel 1982) revealed significant linear ($F(1, 75) = 14.99$, $p < 0.001$; $R^2 = 0.17$) and quadratic ($F(1, 75) = 12.56$, $p < 0.001$; $R^2 = 0.14$) trends; moreover, the drop in influence from 4- to 9-person groups

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mean Levels of Influence and Performance in Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>Source size</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Influence $^a$</td>
<td>0.39 (0.22)$^b$, $^c$</td>
</tr>
<tr>
<td>Performance (MAD)$^h$</td>
<td></td>
</tr>
<tr>
<td>Initial</td>
<td>10.32 (7.59)</td>
</tr>
<tr>
<td>Source$^i$</td>
<td>9.16 (2.39)</td>
</tr>
<tr>
<td>Final</td>
<td>7.88 (4.65)</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td>18 (19)</td>
</tr>
<tr>
<td>Ego-neutral$^i$</td>
<td>7.72 (4.05)</td>
</tr>
</tbody>
</table>

Notes. $n = 20$. Standard deviations are in parentheses.

$^a$Lower MADs indicate better performance.

$^b$Ego-neutral weights were set at 0.50, 0.67, 0.80, and 0.90 for the 1-, 2-, 4-, and 9-person sources, respectively.

$^c$- $^i$Unequal variances by a Brown-Forsythe test.

$^d$- $^g$Shared superscripts indicate significant differences (Tukey HSD tests, $a_{FWS} = 0.05$).
was not significant ($F(1,75) = 0.27$, ns). Thus, the hypothesis of a monotonic increase in informational influence with larger sources (H1) was supported (Braver and Sheets 1993). Groups were more influential than individuals, and larger groups more influential than smaller groups at a declining rate.

Although clearly influenced by the additional judgment, 67 of 80 participants fell short of the ego-neutral level of influence. Differences between ego-neutral and observed levels of influence were calculated for each participant and subjected to a one-way analysis of covariance. Across conditions the average difference was 0.18 ($SD = 0.19$), which was significantly greater than zero ($F(1,75) = 50.69, p < 0.001$). These differences also increased with source size as expected ($F(1,75) = 11.48, MSE = 0.028, p < 0.01; R^2 = 0.13$ for the linear trend). Thus, H2 was supported.

**Performance.** Because participants relied to some extent on the source estimate, they improved their judgments. Final MADs were lower than initial MADs for 73 of 80 participants, with an average improvement of 30% ($SD = 20$%). As Table 1 makes clear, however, improvement was not monotonically greater for larger groups.

Participants underperformed the ego-neutral judgment policies. Final MADs were lower than ego-neutral MADs for only 30 of 80 participants. A repeated-measures analysis of covariance indicated that final MADs were greater than ego-neutral MADs ($F(1,76) = 9.09, MSE = 0.926, p < 0.01; R^2 = 0.11$) and that this underperformance increased with source size ($F(1,75) = 3.69, MSE = 1.65, p < 0.10; R^2 = 0.05$ for the linear trend). In other words, underperformance was trivial for those advised by sole individuals ($d = 0.04$) but substantial for those advised by 9-person groups ($d = 0.98$). This supported H4.

**3.2.4. Discussion.** Even in the absence of contemporaneous feedback, participants made more use of group judgments than those of individual advisors. Participants improved upon their initial judgments but were conservative in their use of advice—increase in influence declined at a faster rate than recommended by an ego-neutral judgment policy. Accordingly, their observed policies failed to weight appropriately the more accurate and reliable judgments. The consequences of this for performance depended on the size (and corresponding accuracy) of the influence source.

Participants in this experiment were modestly sensitive to group size. However, because the size of an influence source and its validity are correlated, it cannot be ruled out that people were influenced by expertise rather than source size. As previously discussed, people are capable of detecting differences in expertise and placing, though imperfectly, more weight on the perceived expert. Despite the lack of explicit feedback
to those of a matched participant in the base condition. Participants in these two conditions also indicated how confident they were following each initial estimate (i.e., not at all, somewhat, fairly, or very confident). Finally, participants in the focal condition were also matched to a participant in the base condition. These participants did not make their own initial estimates. For each city, which was anonymous (e.g., city 1, city 2, etc.), they were provided a focal judgment. This was the initial estimate of their matched participant and his or her confidence in that estimate. After a five-second delay, the average estimate of the 3- or 15-person groups was presented, and participants then made their own estimate.

As in the prior study, participants did not receive feedback during the task, nor could they return to prior estimates. Each city was worth 20 cents, and for each degree Fahrenheit by which a participant’s estimate missed the truth, he or she lost 1 cent. Thus, participants could earn a $4.00 bonus for perfect estimates and were equally penalized for over- or under-estimation. The minimum bonus over all 20 cities was truncated at zero. Participants answered a few questions upon completion of the task, after which they were debriefed and paid.  

### 3.3.2. Results

**Preliminary Analysis.** Table 2 indicates that the expertise of participants (i.e., initial MADs) in the base and control conditions did not vary ($F(3, 48) = 0.17$, $MSE = 23.47$, ns). Nevertheless, because initial MAD was correlated with both influence ($r = 0.27$) and final MAD ($r = 0.40$), it was included as a covariate in the analyses of these measures to increase power. As expected, the validity of the group increased with group size (i.e., source MADs declined with size) in accordance with the law of large numbers ($F(1, 48) = 19.15$, $MSE = 2.63$, $p < 0.001$; $R^2 = 0.29$).

**Influence.** The mean level of influence was 0.60 ($SD = 0.17$), which varied by condition ($F(5, 95) = 7.97$, $MSE = 0.024$, $p < 0.001$; $R^2 = 0.30$). Replicating the pattern of the prior study, participants in the base condition were more influenced by larger groups ($F(1, 47) = 6.82$, $MSE = 0.024$, $p < 0.05$; $R^2 = 0.13$), supporting H1. Forty-six of 50 fell short of using ego-neutral weights with the group estimates, but egocentrism was no greater with the 15-person groups ($F(1, 48) = 1.11$, $MSE = 0.014$, ns).

Two tests were carried out to examine whether sensitivity to the accuracy of the group estimates accounted for the higher influence of 15-person groups in the base condition rather than information about group size. First, if people were sensitive to the accuracy of the additional estimates apart from information about their source, then influence by 15-person groups should have been higher than influence by 3-person groups in the control condition (which had no information about the source of the additional estimate). Although influence by anonymous 15-person groups was slightly higher than that of anonymous 3-person groups (see Table 2), this difference was not significant ($F(1, 47) = 2.43$, $MSE = 0.020$, ns). Second, influence should have been similar in the base and control conditions if perceived accuracy was the primary driver of influence. This was also not the case: Overall influence was significantly greater in the base condition than in the control condition ($F(1, 47) = 15.09$, $MSE = 0.022$, $p < 0.001$; $R^2 = 0.24$). Thus, it appears that information about the groups’ size and not the perceived accuracy of their estimates was primarily responsible for the greater influence of larger groups.

To test the focalism hypothesis, influence in the base and focal conditions was compared. If asymmetric attention to a focal judgment explains the discounting of additional information, then influence should not have differed between those revising their own judgments and those revising identical focal judgments provided to them. Here an interaction emerged with

---

Table 2: Mean Levels of Influence and Performance in Study 3

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>3-person groups</th>
<th>15-person groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td>Focal</td>
<td>Base</td>
</tr>
<tr>
<td></td>
<td>Focal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.16)</td>
</tr>
<tr>
<td>Performance (MAD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial</td>
<td>10.15</td>
<td>11.75</td>
</tr>
<tr>
<td></td>
<td>(3.56)</td>
<td>(6.89)</td>
</tr>
<tr>
<td>Source</td>
<td>6.85</td>
<td>8.48</td>
</tr>
<tr>
<td></td>
<td>(1.97)</td>
<td>(1.77)</td>
</tr>
<tr>
<td>Final</td>
<td>7.03</td>
<td>8.48</td>
</tr>
<tr>
<td></td>
<td>(1.85)</td>
<td>(1.74)</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td>30</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>(17)</td>
<td>(25)</td>
</tr>
<tr>
<td>Ego-neutral</td>
<td>6.05</td>
<td>4.91</td>
</tr>
<tr>
<td></td>
<td>(1.64)</td>
<td>(0.97)</td>
</tr>
</tbody>
</table>

**Notes.** $n = 25$. Standard deviations are in parentheses.

*Ego-neutral weights were set at 0.75 and 0.94 for the 3- and 15-person groups, respectively.
4Lower MADs indicate better performance.
5Unequal variances by a Brown-Forsythe test.

---

5 One of these questions, posed prior to any feedback about performance, asked participants to speculate about the similarity or dissimilarity of the individual temperature estimates constituting the group’s average judgment (7-point Likert scale: $-3 = $Very dissimilar, $+3 = $Very similar). An anonymous reviewer pointed out that participants might (justifiably) place less weight on the average if they believed the individual constituent judgments redundant. There was no association between this rating and influence ($r = -0.01$). Thus, there did not appear to be a relationship between influence and the expected similarity of group members’ judgments.
group size ($F(1, 48) = 8.40, MSE = 0.023, p < 0.01; R^2 = 0.15$). Differences in influence between the base and focal conditions were not significant for the 15-person groups ($F(1, 24) = 0.64, MSE = 0.025, ns$) but were significant for the 3-person groups ($F(1, 24) = 11.63, MSE = 0.021, p < 0.01; R^2 = 0.33$). In the latter case, participants in the base condition made a much greater weight of the group's estimates than those in the base condition (see Table 2). However, because the results were ambiguous, additional analysis was carried out at the item level. If people overweight a focal hypothesis, then the frequency with which an initial (i.e., focal) estimate is chosen as the final estimate should have been similar in the base and focal conditions. Of the 1,924 cases in which the focal and group estimates differed in these conditions, participants in the base condition chose the focal estimate in 17% of cases, whereas those in the focal condition chose it in only 4% of cases ($X^2(1) = 25.98, p < 0.001$). This difference was similar for 3-person (15% versus 5%) and 15-person (20% versus 4%) groups. In other words, people who revised their own estimates were more likely to stick with their initial beliefs than those who provided the initial belief as a focal hypothesis.

Performance. Final MADs were lower for participants advised by 15-person versus 3-person groups ($F(1, 95) = 7.72, MSE = 1.98, p < 0.01; R^2 = 0.08$) and for those in the base, control, and focal conditions ($F(2, 95) = 4.75, p < 0.05; R^2 = 0.09$). Pairwise comparisons of adjusted means indicated that participants in the control condition performed significantly worse than those in the base condition; other differences were not significant (Tukey HSD tests, $\alpha_{tu} = 0.05$).

The performance of participants in the base and focal conditions was compared to that of an ego-neutral judgment policy in which the groups were weighted commensurate with their size (that is, 0.75 and 0.94 for the 3- and 15-person groups, respectively). Analysis of covariance revealed that participants’ intuitive policies were significantly worse than the ego-neutral policy ($F(1, 96) = 41.56, MSE = 1.09, p < 0.001; R^2 = 0.30$) and that this underperformance was greater for those advised by 15-person groups ($F(1, 96) = 2.78, p < 0.10; R^2 = 0.03$). Underperformance did not differ between the base and focal participants.

3.3.3. Discussion. Participants in the base condition replicated many of the findings in Study 2. Influence was higher by the 15-person groups yet fell short of the ego-neutral levels. Performance improved because of this influence and was greater for those advised by 15-person (37%) than 3-person (30%) groups. Nonetheless, because these participants placed too much weight on their initial beliefs, they underperformed the ego-neutral judgment policies.

This study also ruled out an alternative explanation for the influence of group size—that sensitivity to expertise leads people to place more weight on group judgments. If true, then participants in the control condition, who had no information about the size or source of the additional estimates, should have shown a similar pattern of influence as those in the base condition. This was not the case, nor did those in the control condition meaningfully discriminate between judgments provided by anonymous 3- and 15-person groups.

Focalism was offered as one explanation for the egocentric discounting of additional information. In this account, one’s initial belief serves as a focal hypothesis tested against an alternative hypothesis (in this case, the group estimate). If focal hypotheses in general receive undue attention, then it should not matter whether they are proposed by the self or another. The comparison of influence in the base and focal conditions was somewhat equivocal. Although no difference was found at the aggregate level for those advised by 15-person groups, supporting the focalism account, group influence was significantly higher in the focal condition for those advised by 3-person groups, undermining focalism as an explanation. At the item level, however, it was clear that people who made their own estimates were far more likely to prefer them over the alternative, group estimates than those who were simply provided the identical estimates as a focal hypothesis. Overall, the evidence suggests that the discounting of group advice is worse for those revising their own beliefs. Additional explanations for this are left for the general discussion.

Participants in Studies 2 and 3 were insufficiently sensitive to size differences among groups, falling short of the normative, ego-neutral weights in a supermajority of cases. To be fair, group size is difficult to evaluate in isolation. It is not an inherently meaningful attribute, and this lack of evaluability may lead to magnitude insensitivity (Hsee et al. 2005). If people have correct intuitions about the wisdom of crowds (as suggested by Study 1), then direct, joint comparisons of multiple groups should increase the influence of larger ones beyond that seen in the between-participants designs of these studies (Hsee et al. 1999). Study 4 tests this by manipulating group size within participants.

A second manipulation in Study 4 is whether the group judgment is framed as a “consensus” or “average” estimate. Perhaps holding the participants back from placing more weight on the group judgment is a distrust of averaging. Larrick and Soll (2006), for instance, asked students to estimate the accuracy of alternative forecasting strategies. Included in these were a group’s judgment arrived at through discussion and by deference to the most confident group
member, both of which were rated as superior to a simple average of the members’ forecasts. They concluded that the benefits of averaging are apparent to people only in limited circumstances. If true, then group estimates framed as consensus judgments should receive more weight than those framed as averages.

3.4. Study 4

3.4.1. Task and Procedure. One hundred and twenty-five students from a public U.S. university (66 female; M_{age} = 19.8 years) participated in the experiment in exchange for a participation fee of $6 and a bonus for accuracy (M = $0.82, SD = $0.28). The task was to estimate the ages of 38 celebrities aided by a group judgment. The size of the group—2 or 9 people—was manipulated within participants. Participants were randomly assigned to one of four between-participants conditions formed by crossing judgment frame (consensus versus average) and order (2-then-9 versus 9-then-2). Celebrities were randomly chosen from a list of actors and actresses worldwide. Forty-eight separate students estimated their ages on a prior occasion in exchange for a small payment. Groups of eight separate students estimated their ages on a prior occasion. Explanations “the average [consensus] estimate of 2 [9] other students” surveyed on a prior occasion. Questions upon completion of the task, after which they were debriefed and paid.

3.4.2. Results.

Preliminary analysis. The expertise of participants did not vary by judgment frame or order (F(3, 92) = 0.09, MSE = 11.57, ns). Because initial MAD was correlated with both influence (r = 0.37) and final MAD (r = 0.71), it was included as a covariate in analyses of these measures. As expected, 9-person groups were more accurate than 2-person groups (F(1, 58) = 49.63, MSE = 0.799, p < 0.001; R^2 = 0.46).

Influence. Mean influence across the conditions was 0.64 (SD = 0.20). Table 3 summarizes influence across conditions. A repeated-measures analysis of covariance failed to find a main effect of judgment frame (F(1, 92) = 0.68, MSE = 0.034, ns) nor a meaningful interaction of frame with the other factors. There were significant effects of order (F(1, 92) = 6.87, p < 0.05; R^2 = 0.07) and group size (F(1, 120) = 70.03, MSE = 0.016, p < 0.001; R^2 = 0.37), qualified by an interaction between them (F(1, 120) = 14.16, p < 0.001; R^2 = 0.11).

Figure 1 illustrates the interaction of these factors, collapsed across judgment frame. In the first block of revisions (a between-participants comparison), 2-person groups were less influential than 9-person groups with a modest effect size (d = 0.54). In the

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Mean Influence by Condition for Study 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group size^a</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>2-then-9 order (n = 31)</td>
<td>0.59 (0.16)</td>
</tr>
<tr>
<td>9-then-2 order (n = 30)</td>
<td>0.51 (0.19)</td>
</tr>
<tr>
<td>Consensus</td>
<td></td>
</tr>
<tr>
<td>2-then-9 order (n = 32)</td>
<td>0.66 (0.19)</td>
</tr>
<tr>
<td>9-then-2 order (n = 32)</td>
<td>0.49 (0.22)</td>
</tr>
</tbody>
</table>

Note. Standard deviations are in parentheses.

^aRepeated measure. Ego-neutral levels of influence were set at 0.67 and 0.90 for the 2- and 9-person groups, respectively.
second block of revisions, 2-person groups were also less influential than 9-person groups, but the effect size more than doubled ($d = 1.38$). As predicted, joint evaluation of group size led to greater discrimination between them. However, the effect of the joint evaluation was not to increase the influence of 9-person groups as intended, which did not differ whether encountered first or second, but to decrease the influence of 2-person groups.

**Performance.** Only group size was a significant predictor of final MAD ($F(1, 120) = 33.80, \text{MSE} = 0.568, p < 0.001; R^2 = 0.22$). Table 4 summarizes performance, collapsed across order and judgment frame. Final MADs were lower than initial MADs in 92% of cases with 2-person groups and 99% of cases with 9-person groups, with an average improvement of 27% ($SD = 17\%$).

A repeated-measures analysis of covariance revealed that participants continued to underperform the ego-neutral judgment policies ($F(1, 362) = 10.43, \text{MSE} = 0.642, p < 0.01; R^2 = 0.03$), and this underperformance was greater with the 9-person groups ($F(1, 362) = 19.88, p < 0.001; R^2 = 0.05$). Final MADs were similar to those of the ego-neutral estimates for 2-person groups ($F(1, 91) = 1.52, \text{MSE} = 0.757, \text{ns}; d = 0.06$) but were significantly worse for 9-person groups ($F(1, 91) = 17.13, \text{MSE} = 0.494, p < 0.001; d = 0.64$).

**3.4.3. Discussion.** Observed judgment policies were unaffected by the framing of the group judgment as a consensus or average estimate despite research showing that people have more faith in the former. Increasing the evaluability of group size did affect information use, but not as expected. The difference in influence between 2- and 9-person groups was substantially larger in the second block of revisions. Yet this was due not to the enhanced influence of 9-person groups but to the diminished influence of 2-person groups. Use of the former’s judgments was higher than in Study 2 (0.71 versus 0.60, perhaps due to differences in stimuli), which appeared to be a ceiling on the amount of influence by 9-person groups in this task, even when evaluated after a 2-person group. In contrast, evaluating 2-person groups after 9-person groups seemed to provide participants justification for relying less on them. Thus, the enhanced evaluability of group size undermined informational influence instead of bolstering it.

**4. General Discussion**

Optimistic and pessimistic conclusions emerge from this examination of the informational influence of groups. Under conditions free of normative influence, people successfully used information provided by others to improve their judgments. In their behavior, they made monotonically greater use of group judgments, which improved the accuracy of their updated estimates. And in their explicit beliefs they correctly recognized the ordinal relationship between group size and validity on truth-plus-error judgments.

In this study, ego-neutral weighting was prescribed as a straightforward judgment policy less stringent in its demands for rationality than a regression-based approach. Yet even by this more liberal standard, people were suboptimal users of group judgments. This is illustrated in Figure 2, which plots the relationship between group size and influence across four studies.\(^6\)

The graph includes a line for the ego-neutral weights and a fitted regression line for the 11 observed, mean levels of influence in these studies. Whereas the ego-neutral model of influence is described by

\[ \hat{y} = k(\text{group size})^d(\text{group size} + 1)^{-1} \]

with $k = 1$,

the fitted regression line ($R^2 = 0.79$) is described by

\[ \hat{y} = 0.83(\text{group size})^{0.89}(\text{group size} + 1)^{-0.95}. \]

\(^6\) Includes the results of an unreported experiment identical to Study 2 (Study 2A) except that participants were advised by 1-, 4-, or 40-person sources. Mean influence across these conditions was 0.48 ($SD = 0.24$), 0.57 ($SD = 0.12$), and 0.60 ($SD = 0.12$), respectively. Full details are available from the author.
The equations are similar, though even with this small sample their equivalence could be formally rejected ($F(3, 8) = 36.54, p < 0.001$). Group accuracy does exhibit diminishing returns with size, but not at the rate implied by the intuitive judgment policies observed in the present study.7

The present study adds two insights to research on the suboptimal use of additional information in belief revision. First, a lack of sophisticated appreciation for the law of large numbers makes it particularly acute when faced with group judgments. It is one thing to recognize that mean judgments will be more valid with increases in group size; it is another to appreciate that the marginal contribution of an additional opinion to a group, including one’s own, approaches zero quite rapidly as the group increases. Second, the consequences of a suboptimal judgment policy are highly contingent. Participants advised by a sole individual or even by 2-person groups did as well statistically as the ego-neutral policy. Their initial beliefs meaningfully contributed to the final estimates, which were robust to deviations from the ego-neutral policies. Those advised by larger groups, however, substantially underperformed the ego-neutral policies. In these cases, deviations from unit weighting were not robust to the increasing asymmetry in expertise between the participant and group.

Study 3 provided evidence that focalism (Schkade and Kahneman 1998, Wilson et al. 2000) does not explain egocentric judgment policies. Those revising their own judgments were less willing to abandon them than those provided the identical judgments as a focal hypothesis. Three additional explanations are worth exploring in future research. Yaniv and Kleinberger (2000) suggested that the availability of reasons for one’s own beliefs but not another’s increases support for the former (Tversky and Koehler 1994). It is also possible that a more basic attachment to one’s own judgments is responsible (cf. Thaler 2000), regardless of the support for those judgments. Finally, something as fundamental as anchoring or primacy effects may lead to conservatism in the use of additional information (Edwards 1982, Hogarth and Einhorn 1992). Moreover, these are not disjoint theories, and all may contribute to egocentrism in opinion revision.

Although life does provide instances where policy makers make decisions based on aggregate judgments (e.g., average or consensus forecasts of inflation rates, gross domestic product, earnings, etc.), a reasonable objection to these findings is that we are often faced with a distribution of group members’ judgments instead of an aggregate. In the example provided earlier, the sales executive may prefer individual forecasts from the analyst team instead of their consensus or mean forecast. Wilder (1977) argued that influence is more a function of the perceived number of distinct social entities than of the total size of the influence source. He found, for example, that two groups of two were more influential than one group of four. Thus, the more group members are perceived as distinct, the more their collective size has an impact. On the other hand, presenting the distribution of estimates may lead people to pick-and-choose judgments similar to their own. Yaniv and Milyavsky (2007) found that participants faced with a distribution of two, four, or eight opinions “egocentrically trimmed” those discrepant with their own before combining them, which suggests that deemphasizing a group’s collective wisdom encourages people to incorporate only those individual judgments that are redundant with their own.

The effect of knowing the distribution of members’ judgments on influence was therefore examined in an additional study. Fifty-seven participants (21 female; $M_{\text{age}} = 21.8$ years) revised 35 estimates of celebrity ages advised by a 9-person group and were rewarded for accuracy. They were randomly assigned to three conditions: 18 received only the average judgment of the group; 20 received only the individual judgments constituting that average; and 19 received both the average and individual judgments. Mean levels of influence across these conditions were 0.70 ($SD = 0.15$), 0.75 ($SD = 0.30$), and 0.68 ($SD = 0.21$), respectively, which were statistically equivalent ($F(2, 34) = 0.14, MSE = 0.051, ns$) and similar to the level obtained in Study 4. Hence this initial evidence did not indicate that having simultaneous access to the distribution of group members’ judgments has a meaningful effect on the use of their collective wisdom.

5. Practical Implications

Teams are playing an increasing role in organizations (Devine et al. 1999), presumably for their enhanced decision-making ability, among others. It is therefore important to understand the influence of their judgments on decision makers. This research provides preliminary evidence that collective wisdom is underappreciated. Perhaps experience with teams provides ample counterexamples to their benefits: conformity due to normative pressures, the punishment of dissent, unproductive and interpersonal conflict, and so forth. Yet these pitfalls associated with consensus-based decision making can often be mitigated by simply pooling member judgments, as was...
done in this study. If this were practiced more frequently, decision makers could learn that the accuracy of a group’s judgment alone is often superior to the error-prone revision of their own beliefs.

An archetypal example of this practice is the use of prediction or information markets. These allow independent individuals to place small bets on the occurrence of political, economic, or sporting events (e.g., the Arizona Cardinals will win Super Bowl XLIII), where the market value of the trade reflects the collective assessment of the probability of the event’s occurrence. Organizations could use internal markets to help forecast product success, competitor actions, project timelines, and so forth. The lack of appreciation for this collective wisdom documented herein may be one reason information markets have failed to catch on inside organizations (The Economist 2005).

6. Conclusion

Milgram (1965) once commented that “Edifying effects of the group, although acknowledged, have rarely been demonstrated with the clarity and force of its destructive potential” (p. 127). Groups and teams have a mixed reputation in organizations in part due to these demonstrations. Yet he also suggested that “constructive conformity” to a group can benefit individuals, especially if the influence is informational. This is increasingly recognized in principle, but the influence of group judgments on behavior still indicates a shallow appreciation for the wisdom of crowds.

Acknowledgments

The author acknowledges the financial support of Duke University and the invaluable advice, comments, and assistance of Andrew Carton, Lisa Cavanaugh, Kaylen Francis, Abhijit Guha, Richard Larrick, Jack Soll, Amber Turner, John Vickery, and several anonymous reviewers.

References


