Smaller Crowds Outperform Larger Crowds and Individuals in Realistic Task Conditions

Mirta Galesic, Daniel Barkoci, and Konstantinos Katsikopoulos


CITATION
Smaller Crowds Outperform Larger Crowds and Individuals in Realistic Task Conditions

Mirta Galesic
Santa Fe Institute, Santa Fe, New Mexico, and Max Planck Institute for Human Development, Berlin, Germany

Daniel Barkoczi and Konstantinos Katsikopoulos
Max Planck Institute for Human Development, Berlin, Germany

Decisions about political, economic, legal, and health issues are often made by simple majority voting in groups that rarely exceed 30–40 members and are typically much smaller. Given that wisdom is usually attributed to large crowds, shouldn’t committees be larger? In many real-life situations, expert groups encounter a number of different tasks. Most are easy, with average individual accuracy being above chance, but some are surprisingly difficult, with most group members being wrong. Examples are elections with surprising outcomes, sudden turns in financial trends, or tricky knowledge questions. Most of the time, groups cannot predict in advance whether the next task will be easy or difficult. We show that under these circumstances moderately sized groups, whose members are selected randomly from a larger crowd, can achieve higher average accuracy across all tasks than either larger groups or individuals. This happens because an increase in group size can lead to a decrease in group accuracy for difficult tasks that is larger than the corresponding increase in accuracy for easy tasks. We derive this nonmonotonic relationship between group size and accuracy from the Condorcet jury theorem and use simulations and further analyses to show that it holds under a variety of assumptions. We further show that situations favoring moderately sized groups occur in a variety of real-life situations including political, medical, and financial decisions and general knowledge tests. These results have implications for the design of decision-making bodies at all levels of policy.

Keywords: wisdom of crowds, majority rule, Condorcet jury theorem, group decision making

Supplemental materials: http://dx.doi.org/10.1037/dec0000059.supp

Individuals and societies often make decisions by following the majority vote of moderately sized groups. For example, jury sizes in many countries range from six to 15 people who most often decide by simple majority (Leib, 2008). Local town and parish councils such as those in the United Kingdom and Australia consist of five to around 30 members (Electoral Council of Australia & New Zealand, 2013; U.K. Department for Communities and Local Government, 2008); governing bodies of most German labor unions have from three to 35 members (dejure.org, 2013); parliamentary committees in the United States, the European Union, Australia, and other countries have on average 20 to 40 members (European Parliament, 2014; Haas, 2014; Parliament of Australia, 2014); subcommittees in the U.S. House and Senate consist of on average 10 to 15 people (Haas, 2014); and policy boards of most central banks have up to 12 members (Lybek & Morris, 2004). Similarly, individuals considering a va-
riety of decisions typically rely on six or fewer close friends (Galesic, Olsson, & Rieskamp, 2012) and read about five and rarely more than 30 online reviews before deciding whether to trust a business (Anderson, 2014). Deciding in moderately sized groups can also be observed in other species throughout the animal kingdom (Krause & Ruxton, 2002).

In many cases, deciding in groups rather than relying on an individual decision maker can boost overall decision accuracy (Surowiecki, 2004). This has been shown both for predictions of continuous variables, such as in Galton’s demonstrations of the value of vox populi (Galton, 1907), and for categorical choices between distinct courses of action under certain conditions (Condorcet, 1785). Today, technological advances make communication and meeting in larger groups easier than ever before (e.g., various social networking sites; LiquidFeedback, 2014). Why, then, do most committees remain moderately sized, and why do most people consult only a limited number of others’ opinions? Existing explanations focus on time and coordination costs or on cognitive limitations that prevent stable relationships with a large number of individuals (Dunbar, 1993). Here, we put forward an argument for the superiority of moderate group sizes based solely on group decision accuracy.

Committees are often formed of experts in a particular domain. Consequently, on most of the tasks they encounter average individual accuracy of group members will be above chance. However, every now and then committees might encounter a surprisingly difficult task on which most group members will be wrong. We show that in real-life situations average group accuracy across different tasks often peaks when groups are moderately sized. This occurs without any selective sampling of group members on the basis of expertise (Budescu & Chen, 2015; Goldstein, McAfee, & Suri, 2014; Mannes, Soll, & Larrick, 2014). Rather, group members can be selected randomly from a larger crowd. The advantage of moderately sized groups occurs solely because the accuracy of groups deciding by simple majority or plurality rules increases with their size for relatively easy tasks and at the same time decreases for tasks for which most individuals make the wrong prediction.

Tasks With Surprising Outcomes

Tasks with surprising outcomes that are difficult to predict can be found in many domains, including political and economic forecasts, medical diagnoses, and general knowledge tests. Consider election forecasts. Expert forecasters often show better-than-chance prediction accuracy, but a few election years have been surprisingly difficult to predict. Such was the 2015 general election in the United Kingdom, where all but one polling company erroneously predicted that the Tories would not win a majority of seats in the Parliament (Bialik, 2015). Similarly, the majority of forecasters in the 2000 presidential elections in the United States predicted Gore’s victory over Bush in Florida (Graefe, 2014; Whitson, 2001).

In such tasks, the majority of individuals can be wrong, resulting in average individual accuracy below 50% on those particular tasks. This can happen when tasks are characterized by the so-called Brunswikian uncertainty (Juslin & Olsson, 1997) that occurs because of imperfect correlations of environmental cues and the actual states of the world they are used to predict. If most people rely on the same cues (or, equivalently, opinion leaders, media reports, etc.) to make inferences, cases where a cue (leader, report) is misleading can create situations where a majority of people are incorrect. Consider the knowledge question “Which city is farther north, New York or Rome?” which most people answer incorrectly. Temperature, a cue that is valid for most other comparisons of city latitudes, points to the wrong answer for this pair of cities (Gigerenzer, Hoffrage, & Kleinbölting, 1991).

Even when individuals rely on different cues, these cues could all fail to predict the correct outcome for some specific tasks, either because they are not suited for predicting some particular cases or because the environment changed between the moment of prediction and the moment when the outcome was observed. For instance, most diseases might be accurately diagnosed on the basis of their symptoms, but some less well known or rare diseases have symptoms that can point to several different diagnoses; or forecasts of economic growth may prove to be wrong in some cases because of unobservable underlying complexities affecting the financial markets. In what follows, we call tasks with
surprising outcomes that an average expert predicts incorrectly "difficult" and those an average expert predicts correctly "easy."

Group Size and Accuracy Over a Range of Task Difficulties

Most committees face a variety of task difficulties, ranging from very easy to quite difficult, in the course of their existence. However, most past studies of group decision accuracy have typically investigated situations in which groups encounter tasks of the same and known difficulty (e.g., Grofman, Owen, & Feld, 1983; List & Goodin, 2001). Once task difficulty is known, it is relatively straightforward to tell what the group size should be to maximize accuracy, at least when group members vote independently. In principle, for easy tasks, in which average individual accuracy of group members (average individual probability of being correct) is larger than 0.5, majority vote in larger groups will be more accurate than in smaller groups, and vice versa for difficult tasks (Condorcet, 1785).

In most real-life situations, however, one cannot predict in advance how difficult the next task will be. All one might know is an approximate distribution of task difficulties a group might face. For instance, an expert group might encounter mostly easy tasks and occasionally some surprisingly difficult tasks. A novice group might find most tasks difficult and some easy. In addition, in some domains predictions are inherently easier than in others. Not knowing exactly what task difficulties a group will face, can one say anything about the group size that will lead to highest achievable accuracy?

Wisdom of Small, Randomly Selected Crowds

Here we show that in many plausible real-life situations moderately sized groups will achieve higher accuracy than either larger groups or individuals. We focus on tasks in which groups need to vote for one of two or more possible courses of action and decide by simple majority. We assume that it is eventually possible to determine whether the group decision was correct. The voting stage may or may not be preceded by a group discussion in which members determine common ground for understanding the problem, share some or all of the information they possess individually, make various continuous judgments relevant for the problem, and discuss consequences of taking one or the other course of action. We focus on the final stage of the decision-making process, in which individual votes are transformed into a group vote for one of two or more possible courses of action. In this stage, majority and plurality rules can perform as well as or better than can computationally more demanding decision procedures (Hastie & Kameda, 2005).

For simplicity, we first analyze situations with tasks involving two options between which groups choose by simple majority rule, where there are only two task difficulties, assuming that individual group members vote independently and assuming that they are selected from a large population with replacement. Afterward, we add a number of more realistic assumptions, allowing for more than two task difficulties and more than two options in each task, for correlated judgments of group members and for sampling of group members from a finite group without replacement. Finally, we provide several real-world examples from different task domains in which smaller groups can perform better than larger ones can.

Two Task Difficulties

To determine how group accuracy depends on group size when a single task involves making a choice between two options using a simple majority rule, one can straightforwardly use the Condorcet jury theorem (CJT). It can be written formally as a binomial cumulative distribution function:

$$P_n = \sum_{i=m}^{n} \binom{n}{i} p^i (1-p)^{n-i},$$  \hspace{1cm} (1)

where $P_n$ is group accuracy at group size $n$, $m$ is size of simple majority, and $p$ is average individual accuracy. Without loss of generality, $n$ is assumed to be always odd. Other voting rules are possible, such as requiring two thirds majority or unanimous decision, but it has been shown that simple majority leads to best performance in noisy environments (Sorkin, West, & Robinson, 1998). Individual group members
can have heterogeneous skills. As long as the distribution of individual skills is symmetrical, CJT predictions remain essentially the same as if all individuals had the same skill level (Grofman, Owen, & Feld, 1983). With increase in $n$, group accuracy $P_n$ monotonically increases to 1 for tasks with average individual accuracies $P > 0.5$ and monotonically decreases to 0 for tasks with $P < 0.5$. Deviations occur only in exceptional cases, for instance when some individuals consistently have accuracy 0 or 1 and groups are very small; but even in these cases the average of individual accuracies will converge to a value close to 0.5 (Owen, Grofman, & Feld, 1989). In other words, CJT predictions generalize to a large range of asymmetrical distributions of individual skills (Grofman et al., 1983).

To study average group accuracy over two or more tasks, we first assumed that groups encounter two task difficulties: With probability $e$ they encounter easy (denoted E) tasks, for which average individual accuracy $P_E > 0.5$, and with probability $1 - e$ they encounter surprising or difficult (denoted D) tasks, for which average individual accuracy $P_D < 0.5$. Figure 1 shows how average group accuracy $\bar{P}$ across the two task difficulties changes with increase in group size $n$, assuming that the proportion of easy tasks is $e = 0.6$. Following CJT (see Equation 1), for easy tasks group accuracy $P_E$ is larger than $\bar{P}$ and increases monotonically to 1 as groups get larger (dashed lines in all panels of Figure 1). For difficult tasks, $P_D < \bar{P}$ and decreases monotonically to 0 with increase in group size (dotted lines). The average group accuracy $\bar{P}$ (full lines) is equal to the average of group accuracies on easy and difficulty tasks, weighted by the proportion of tasks of each difficulty that the group encounters, shown in the following equation:

$$\bar{P}_n = eP_{E,n} + (1 - e)P_{D,n},$$  

(2)

where $P_n$ is the average accuracy of a group of size $n$, $e$ is the proportion of easy tasks, and $P_{E,n}$ ($P_{D,n}$) is the accuracy of a group of size $n$ on easy (difficult) tasks derived by the CJT (see Equation 1).

As Figure 1 illustrates, changes in $\bar{P}$ with changes in $n$ depend on the type of task environment. In a “friendly” task environment, easy tasks are quite easy and difficult tasks are not too difficult. Such a task environment might be

![Figure 1](image-url)
encountered by a group of experts who are skilled in solving particular tasks and, even when surprised, do not do too badly. In contrast, an “unfriendly” task environment might more often be encountered by a group of novices: Here, difficult tasks are very difficult and even easy tasks are not too easy.

More formally, we define a “friendly” environment as one in which \( p_E + p_D > 1 \), a “neutral” environment as one in which \( p_E + p_D = 1 \), and an “unfriendly” environment as one in which \( p_E + p_D < 1 \). These definitions express whether it is the accuracy in easy tasks or the accuracy in difficult tasks that is further away from chance. For example, \( p_E + p_D > 1 \) is equivalent to \( p_E - 0.5 > 0.5 - p_D \), which means that in friendly environments, the accuracy in easy tasks is above chance more than the accuracy in difficult tasks is below chance.

In all environments, \( P_n \) will start from \( P_1 = e p_E + (1 - e) p_D \), which is the average individual accuracy across easy and difficult tasks, and with increase in \( n \) it will eventually converge to the proportion of easy tasks \( e \). Convergence to \( e \) rather than to 0 or 1 as would be predicted by the simple CJT happens because for large enough \( n \), \( P_E \) reaches 1 and \( P_D \) reaches 0, so \( P_n \) converges to \( e \times 1 + (1 - e) \times 0 = e \).

In between these two extremes, \( P_1 \) and \( e \), \( P \) can be a monotonically increasing, monotonically decreasing, U-shaped, or inverted-U-shaped function of \( n \). Which of these shapes is completely determined by two factors defined precisely earlier: the type of environment (friendly, neutral, or unfriendly) and the proportion of easy tasks \( e \).

More precisely, the following holds as \( n \) increases to \( n + 2 \) (the next odd group size):

\[
\Delta P_n = \begin{cases} 
0 & \text{if } \Delta P_{E,n} > \frac{1 - e}{e} \Delta P_{D,n} \\
< 0 & \text{if } \Delta P_{E,n} < \frac{1 - e}{e} \Delta P_{D,n} \\
> 0 & \text{if } \Delta P_{E,n} < \frac{1 - e}{e} \Delta P_{D,n} 
\end{cases}
\]

where \( \Delta P_n = P_{n+2} - P_n \) is change in average group accuracy across all tasks and \( \Delta P_E = |P_{E,n+2} - P_{E,n}| \) and \( \Delta P_D = |P_{D,n+2} - P_{D,n}| \) represent change in average group accuracy across easy and difficult tasks, respectively. In words, average group accuracy \( P \) will increase with group size if the rate of change in accuracy on easy tasks \( \Delta P_{E,n} \) is higher than the rate of change in accuracy on difficult tasks \( \Delta P_{D,n} \), weighted by the relative prevalence of difficult tasks \( (1 - e)/e \). Put more simply, if an increase from \( n \) to \( n + 2 \) leads to a gain in \( P_E \) that is larger than the weighted loss it produces in \( P_D \), \( P \) will increase and otherwise decrease. It will reach its peak when the gains and weighted losses cancel each other.

To illustrate how Equation 3 works, consider a friendly environment, in which \( p_E - 0.5 > p_D \). Here, the rate of change in accuracy on easy tasks is initially higher than the rate of change on difficult tasks, as follows from Equation 1 when \( p_E \) is closer to 1 than \( p_D \) is to 0. However, the initially increasing trend in \( P \) may be reversed as \( n \) continues to increase because \( P_E \) will reach its limiting value, 1, whereas \( P_D \) will still be decreasing toward 1, driving \( P \) down. This is what happens in Figure 1A. In this friendly task environment, an increase in \( n \) initially leads to an increase in \( P \), here peaking at 0.7 for \( n = 7 \) before decreasing to \( e \).

Similar analysis can be applied to less friendly task environments. The component of \( P \), \( e P_E \), or \( (1 - e) P_D \), whichever initially changes faster, will be the first to converge to its limiting value, and then the other component will start changing faster. Figure 1B shows a case of a neutral environment, where \( P \) increases monotonically with \( n \) until it reaches \( e \). Figure 1C shows a particularly interesting case that occurs in unfriendly environments. Here, a downward peak occurs, with \( P \) initially decreasing and then slowly increasing toward \( e \). Note that in this case \( P \) will ultimately become larger than 0.5 (because \( e = 0.6 \)) even though the average individual accuracy across different task difficulties was lower than 0.5 ((\( P_1 = 0.44 \)).

Solving Equation 3 analytically involves taking derivatives of the binomial cumulative distribution functions \( P_E \) and \( P_D \) with respect to \( n \). This produces cumbersome solutions, so approximations have been developed for large \( n \) (Grofman et al., 1983; Marsaglia & Marsaglia, 1990). To examine how changes in small \( n \) relate to group accuracy for different combinations of task difficulties, we calculated \( P \) using Equation 2 across a range of group sizes, for all combinations of easy
(0.6 ≤ \( p_E \) ≤ 0.9) and difficult (0.1 ≤ \( p_D \) ≤ 0.4) tasks, separately for different proportions of easy tasks (0.1 ≤ \( e \) ≤ 0.9), in increments of 0.1. Results presented in Figure 2 and in Figure S1 in the online supplemental materials show that non-monotonic changes in \( P \), such as those shown in Figure 1, occur in more than half of all possible combinations of task difficulties.

Figure 2. Average group accuracy depends on combination of task difficulties and proportion of easy tasks. Each panel shows changes in average group accuracy \( P \) as a function of group size \( n \), different combinations of easy (0.6 ≤ \( p_E \) ≤ 0.9) and difficult (0.1 ≤ \( p_D \) ≤ 0.4) tasks and different proportions of easy tasks (0.1 ≤ \( e \) ≤ 0.9). In each panel, upper lines represent higher proportions of easy tasks \( e \) (see legend to the right of each row). Panels above the diagonal represent friendly task environments, those in the diagonal neutral, and those below the diagonal unfriendly task environments. Circles show the maximum value of \( P \) for each case. Dotted lines in the panels in the upper left corner denote task environments observed in real-world policy tasks. See the online article for the color version of this figure.
Moderate group sizes have the advantage over larger groups, and often also over single individuals, in all friendly environments (\( \bar{p}_E + \bar{p}_D > 1 \), subplots above the diagonal) whenever the proportion of easy tasks is \( e > (0.5 - \bar{p}_D)/(\bar{p}_E - \bar{p}_D) \), that is, when average individual accuracy across tasks, \( \bar{P}_i = e\bar{p}_E + (1 - e)\bar{p}_D \), is larger than chance (\( \bar{P}_i > 0.5 \)). In addition, moderate group sizes are as good as larger group sizes in neutral (\( \bar{p}_E + \bar{p}_D = 1 \), subplots on the diagonal) and unfriendly (\( \bar{p}_E + \bar{p}_D < 1 \), subplots below the diagonal) environments whenever easy tasks are very easy (\( \bar{p}_E = 0.8 \)) and are encountered more than half of the time (\( e > 0.5 \)). In these cases, the group accuracy quickly converges to \( e \) and a further increase in \( n \) does not provide additional improvement.

In unfriendly environments (subplots below the diagonal) in which easy tasks are not too easy (\( \bar{p}_E < = 0.7 \)), best results are achieved by either large groups (especially when proportion of easy tasks \( e \) is high) or by single individuals (when \( e \) is low). In this region, moderately-sized groups do not have an advantage.

In sum, the analysis presented so far shows that small groups can be more accurate than larger groups when expert groups, whose members are more accurate than chance on an average task, encounter mostly quite easy tasks but are sometimes confronted with moderately difficult tasks that a group encounters, one of which is the following extension of Equation 2:

\[
P_n = \frac{1}{T} \sum_{i=1}^{T} P_{i,n}, \tag{4}
\]

where \( T \) is the number of tasks and \( P_{i,n} \) is group accuracy on a given task \( i \) at group size \( n \), calculated using Equation 1.

More generally, instead of assuming that task difficulties \( \bar{p}_E \) and \( \bar{p}_D \) are the same for all easy and difficult tasks that a group encounters, one can model them as random draws from beta distributions with parameters \( \alpha_E = \bar{p}_E c \) and \( \beta_E = (1 - \bar{p}_E) c \) for easy tasks and parameters \( \alpha_D = \bar{p}_D c \) and \( \beta_D = (1 - \bar{p}_D) c \) for difficult tasks, where \( c \) is a constant that determines the proportion of hard tasks is larger than the proportion of easy tasks, group performance will decrease with increasing group size (see Figure 2, in particular panels in the lower right corner). Fourth, we delineate conditions for nonmonotonic trends in group accuracy with downward peaks, that is, when moderate group sizes are less accurate than both single individuals and large groups. Fifth, as mentioned before, we test our findings under a variety of assumptions, including two or more task difficulties, tasks with two and more options, independent and correlated votes, and sampling from either infinite populations or from smaller populations without replacement. Finally, we show that situations favoring moderately sized groups occur in a variety of real-life domains including political, medical, and financial decisions and general knowledge tests.

More Than Two Task Difficulties

So far we have assumed, for simplicity, that a group faces only two task difficulties: the same average individual accuracies \( \bar{p}_E \) and \( \bar{p}_D \) for all easy and difficult tasks, respectively (although on each task individuals could have heterogeneous skills). In real life, groups face tasks of a wide range of difficulties. Average group accuracy across many different tasks can be calculated by taking into account the absolute value of correct decision and the cost of utilizing additional group members; rather it is enough to assume that a correct decision is more valuable than an incorrect one. Third, we disprove the assumption of Grofman et al (p. 355) that whenever the proportion of hard tasks is larger than the proportion of easy tasks, group performance will decrease with increasing group size (see Figure 2, in particular panels in the lower right corner).

---

We thank our reviewers for suggesting to investigate some of these issues.
size of the variance of task difficulties. Then, easy tasks have mean difficulty \(\alpha \beta \) and variance \(\alpha \beta k\) where \(\alpha\) is the mean and \(\beta\) is the variance of task difficulties. Figures S2A–S4B in the online supplemental materials show that main patterns of results just described hold even for a wide range of task difficulties.

**Tasks With More Than Two Options**

What if tasks involve plurality choices between more than two options? CJT can be extended to these situations: Group accuracy will increase with group sizes \(n\) as long as the average individual is more likely to choose the correct option over any other option (List & Goodin, 2001). The probability that a group chooses the correct one of \(k\) options can be calculated as a multinomial probability of all \(k\)-tuples of individual votes for the \(k\) options for which the correct option is the plurality winner, given probabilities \(p_1, p_2, \ldots, p_k\) that an average individual chooses each of the \(k\) options. Once group accuracies \(P_{i,n}\) are calculated in this way for different tasks \(t\) and group sizes \(n\), Equation 4 can be used to calculate average group accuracy. One can then show that nonmonotonic changes in average group accuracy \(\bar{P}\) occur in these situations as well, in accord with conditions discussed earlier.

**Effect of Correlated Votes**

So far we have assumed that group members are independent in the sense that they rely on diverse (or uncorrelated) cues to make their judgments. Surprising outcomes can drive a majority of people in the wrong direction even when individuals vote independently. This can happen if the environment changes in a way that makes all cues incorrect or if by chance uncorrelated cues happen to be wrong on the same task. However, the assumption of perfect independence is unrealistic (see, e.g., Broomell & Budescu, 2009). In real life people are often influenced by the same cues, such as the same pieces of information, media reports, or opinion leaders. It has been shown that the presence of opinion leaders or common information that introduces correlations between individuals’ decisions can reduce or even reverse the positive effects of larger group size on group accuracy (Boland, Proschan, & Tong, 1989; Kao & Couzin, 2014; Spiekermann & Goodin, 2012).

There are many ways to model correlation between group members. A number of models have been developed for studying dependencies between experts’ continuous judgments (e.g., Broomell & Budescu, 2009; Clemen & Winkler, 1985; Hogarth, 1978). Boland et al. (1989; see also Spiekermann & Goodin, 2012) developed an approach for studying dependencies between categorical expert judgments and their effect on groups’ majority-based decisions. In this model, which we use here, correlations occur because some voters follow an opinion leader (who does not vote but influences some group members to decide in a certain way) or a common cue. The effects of such correlations can be parsimoniously incorporated in the present framework. Whenever the leader or the common cue is correct, average individual accuracy improves and the task in effect becomes easier. Conversely, whenever the leader or the common cue is wrong, the average individual becomes less accurate and the task becomes more difficult. Hence, the overall accuracy of a group in which some members follow a leader or a common cue can be represented as an average of the group’s performance on easy and difficult tasks. Accordingly, single peak functions of the kind presented earlier can be expected to occur in some conditions.

More formally, following an opinion leader or voting on the basis of a common cue can be studied as a combination of easy tasks (when the leader or cue is correct) and difficult tasks (when the leader or cue is not correct). More precisely,

\[
\bar{P}_n = l(P_n | \bar{P}(1 - r) + r) + (1 - l)(P_n | \bar{P}(1 - r)) ,
\]

where \(P_n\) is average accuracy of a group of size \(n\) across tasks, \(l\) is probability that an opinion leader is accurate on any task, \(P_n\) is group
accuracy at group size $n$ (here depending on $\bar{p}$ and $r$), $\bar{p}$ is average individual accuracy that group members would have without the opinion leader, and $r$ is the proportion of group members who are following the opinion leader. The higher $r$, the higher the correlation among group members (Spiekermann & Goodin, 2012). It is easy to see that when the leader is accurate the tasks will overall be easier (i.e., group accuracy will be higher) than when the leader is not accurate. Thus Equation 5 resembles Equation 2, and a condition similar to Equation 3 must be satisfied for $P$ to increase with group size $n$:

$$(P_{n+2} | \bar{p}(1-r) + r) - (P_n | \bar{p}(1-r) + r)$$

$$> \frac{1-l}{l} \left| (P_{n+2} | \bar{p}(1-r)) - (P_n | \bar{p}(1-r)) \right|$$

(6)

A similar case can be made for situations in which correlations occur because individuals use the same sources of information.

To further explore effects of correlated votes on decision accuracy across different tasks, we introduce correlations between voters on each task, before averaging across tasks. Following Boland et al. (1989), we assume that on each task a proportion of $r$ voters are following a leader or some other cue that is stochastically correct with probability $l$, and as a result their votes become correlated. Whenever the leader or cue is correct, all $r$ voters are correct, and accuracy of the remaining voters depends on their individual skill. More precisely, Equations 2 and 5 can be combined to account for both correlated votes and task difficulty as follows:

$$\bar{P}_n = e(l_E(P_{E,n} | \bar{p}_E(1-r) + r)$$

$$+ (1-l_E)(P_{E,n} | \bar{p}_E(1-r))]$$

$$+ (1-e)(l_D(P_{D,n} | \bar{p}_D(1-r) + r)$$

$$+ (1-l_D)(P_{D,n} | \bar{p}_D(1-r))),$$

(7)

where meanings of the symbols are like in Equations 5 and 6.

We repeated the analyses presented in Figure 2 (see also Figure S1B in the online supplemental materials) while increasing the assumed proportion of voters $r$ who follow the leader from 0 to 1 in steps of 0.1. With increase in $r$, changes in group accuracy with its size become less and less prominent, and for high $r$ there is almost no change in group accuracy with increase in its size. Because in the real world it is difficult or impossible to know what proportion of people will follow a leader in a particular task, we averaged the results over the whole range of values of $r$. The results, shown in Figure 3 and Figure S5 of the online supplemental materials, demonstrate that in most situations the superiority of moderate group sizes still holds under the assumption of correlated votes. More generally, as Figure 3 shows, the increase in group accuracy with $n$ is much smaller when votes of group members are correlated than when they are independent.

### Sampling of Group Members Without Replacement From a Finitely Sized Population

Modeling group accuracy using Condorcet jury theorem assumes that group members are sampled with replacement from a very large population. However, in real life group members of a smaller committee will typically be selected without replacement from a larger, but finitely sized, committee. For such situations, the hypergeometric distribution is a more appropriate model (Tideman & Plassmann, 2013). Specifically, Equation 1 could be rewritten in the following way, using cumulative hypergeometric rather than cumulative binomial distribution:

$$P_n = \sum_{i=m}^{n} \frac{\binom{I}{i} \binom{N-I}{n-i}}{\binom{N}{n}}$$

(8)

where $P_n$ is group accuracy at group size $n$, $m$ is size of simple majority of group of size $n$, $N$ is population size (or size of the larger committee

---

\[ In these simulations, we assumed that the leader has the same skill as does the average group member ($l_E = p_E$ and $l_D = p_D$). The results still hold if one assumes that the leader is 10 percentage points more or less likely to be accurate than the average member. Results for all combinations of $r$ and $l$ are available from the authors or by running the code in the online supplementary materials. \]
from which the smaller group of experts is randomly selected), $I = N/n$, and $I/N$ equals $\bar{p}$, the average individual accuracy in the population.

The binomial distribution used in CJT is simpler and analytically more tractable and is therefore typically used to analyze voting models (Grofman et al., 1984; List & Goodin, 2001). We adopted it here to enable comparison of our results with those of previous studies, but to check whether any conclusions presented in the article would be different when using hypergeometric distribution, we reran all the analyses using the latter model and assuming different realistic sizes of the finite population. For instance, Figure S6 in the online supplemental materials shows the results assuming that members of smaller groups were randomly selected from a finite population of size $N = 31$, without replacement. All conclusions remained qualitatively the same and if anything were more pronounced under these more realistic assumptions.

**Figure 3.** Group accuracy after replicating the analysis in Figure 2 with correlated votes. The results shown are averaged over correlation levels $r$ ranging from 0 to 1 in steps of 0.1. Leader accuracy $l$ is assumed to be equal to the average individual accuracy $\bar{p}$. See the online article for the color version of this figure.
Real-World Illustrations

What is the best committee size in real-world environments? To answer this question, one needs to have a rough idea of the distribution of task difficulties a typical committee might encounter in the real world. Given that committees are usually composed of people who are experts in the relevant area, one could expect that they are on average more accurate than chance. In addition, one could expect that their accuracy in easy tasks is above chance more than their accuracy in difficult tasks is below chance. In other words, a typical task environment in which committees need to make decisions might more often be friendly than unfriendly.

Studies documenting expert accuracies across a range of tasks in the fields of politics, health, and economics support these expectations. For example, in a longitudinal study of expert forecasters of five U.S. presidential elections, Graefe (2014) found that their average individual accuracy across all years was above chance ($P_1 = 0.66$). In easy years, average individual accuracy $p_E$ was 0.88, and in two difficult years (Bush vs. Gore in 2000 and Bush vs. Kerry in 2004), average individual accuracy $p_D$ was 0.34 (see gray dots in Figure 4A). Similarly, a review of accuracy of medical diagnoses for 11 diseases showed that the average individual doctor’s accuracy was above chance ($P_1 = 0.70$). For diseases that were easy to diagnose, average individual accuracy $p_E$ was 0.81, and for difficult ones, including Lyme disease, pyrogenic spinal infections, and abdominal aortic aneurysms, $p_D$ was 0.41 (Schiff et al., 2009; see gray dots in Figure 4B). Furthermore, a review of accuracy of predictions given by the top officials of the U.S. Federal Reserve Bank about future economic trends showed that their average individual accuracy when predicting whether unemployment, economic growth, and inflation would increase or decrease was rather high ($P_1 = 0.71$). Two of the domains were relatively easy to predict ($p_E = 0.86$), whereas economic growth was somewhat difficult ($p_D = 0.43$; Hilsenrath & Peterson, 2013; see gray dots in Figure 4C). Finally, in a study including 120 general knowledge tasks such as which of two randomly selected cities is farther north or which of two randomly selected countries is larger or more populated, Juslin (1994) found that on average participants were quite accurate ($P_1 = 0.76$) but tended to be incorrect on a subset of tasks in which otherwise useful cues pointed to the wrong answers (see inset in Figure 4D). On easy tasks, average individual accuracy $p_E$ was 0.86, and on the difficult tasks $p_D$ was 0.38.

In all of these examples task environments were friendly ($p_E + p_D > 1$), and each expert had above chance accuracy on an average task ($P_1 > 0.5$). These conditions satisfy the conditions outlined earlier for the situations where groups of moderate sizes are likely to reach the highest accuracy.

If one assumes that a policymaker or an individual needs to decide on the best group size to solve tasks illustrated in Figure 4, what group size would reach the highest accuracy? Given these task environments, how many political experts should a journalist consult to improve election forecasts, how many doctors should a patient consult to improve the accuracy of a medical diagnosis, how many economists should a government consult to make a good guess about the future course of the economy, and how many individuals should one consult to maximize one’s chances of giving a correct answer to a general knowledge question? To investigate this, we used Equation 4 to combine group accuracies for different tasks (gray lines in Figure 4) and get average group accuracy in each of the four domains illustrated earlier (thick black line in Figure 4). This analysis shows that the best group size for improving election forecasts by political experts in this particular illustration is $n = 5$. For diagnosing a variety of health problems, the best size of a panel of medical experts in this example would be $n = 11$. For economic tasks such as those faced by Federal Reserve officials, the best group size seems to be $n = 7$. Perhaps coincidentally, this is the designated number of seats on the Federal Reserve’s Board of Governors, although at the moment of writing this article two of those seven seats are empty (Federal Reserve, 2015). Finally, for answering general knowledge items correctly, the best group size for participants of Juslin’s (1994) study is $n = 15$.

Discussion

Our results suggest that the highest accuracy across a diverse set of tasks involving majority vote between two or more courses of action
Real-world environments are often friendly, and group accuracy peaks at moderate group sizes. Gray dots in Panels A–C: Average individual accuracies for particular tasks (five election forecasts in A, diagnoses for 11 diseases in B, forecasts for three economic trends in C). Inset in Panel D: Histogram of average individual accuracies for 120 knowledge tasks. Gray lines: Group accuracy for different group sizes, for each of the different tasks faced by (Panel A) experts predicting U.S. political elections in years 1992, 2000, 2004, 2008, and 2012 (Graefe, 2014); (Panel B) doctors giving medical diagnoses for a range of diseases (AC = acute cardiac ischemia; BC = breast cancer; S = subarachnoid hemorrhage; D = diabetes; G = glaucoma; St = Soft tissue pathology; C = cerebral aneurysm; Bi = brain and spinal cord biopsies; L = lyme disease; P = pyrogenic spinal infections; AA = abdominal aortic aneurysm; Schiff et al., 2009); (Panel C) U.S. Federal Reserve Bank officials giving economic forecasts about future economic trends in unemployment, inflation, and economic growth (Hilsenrath & Peterson, 2013); and (Panel D) individuals answering 120 general knowledge items about sizes, latitudes, and populations of cities and countries (Juslin, 1994). In Panels A–C each gray line represents one task; in Panel D each gray line depicts several tasks, and frequency of different tasks at each level of task difficulty (p) is shown in the inset. Note that in all domains easy tasks prevail, accompanied with a few surprising tasks that were difficult for most participants. Thick black lines indicate average group accuracy across different tasks. In all four examples, average group accuracy peaks at moderate group sizes (as indicated by circles): in Panel A at n = 5, in Panel B at n = 11, in Panel C at n = 7, and in Panel D at n = 15.
might often be achieved by moderately sized rather than large groups. We provide novel results regarding the precise conditions under which this phenomenon occurs and show that it holds even if one assumes that individuals have diverse skills, that their votes are correlated, that tasks have more than two options, or that groups encounter more than two task difficulties.\footnote{An overview of our main formalisms is given in Table S1 in the online supplemental materials.}

Although best group size depends on individual accuracy on easy and difficult tasks and the proportion of easy tasks, we show that conditions favoring relatively small committees might hold in many real-world situations. In these situations, groups of experts have to decide about a variety of issues over time, among which most are relatively easy to solve but some produce surprising outcomes. Of course, real-world group sizes are influenced by many factors other than accuracy, but our results show that groups that are smaller because of organizational or communication constraints do not necessarily have to be less accurate than larger groups.

Even though the differences in accuracy between groups of moderate and larger sizes might sometimes be small, they are still relevant. The cost of larger groups is almost always larger than that of smaller groups (Libby & Glass, 2010), especially when groups are composed of experts and are engaged over longer time periods. If larger size does not improve group accuracy, there is no reason to include even one additional member.

Note that we modeled tasks in which groups use simple majority or plurality rules to choose between discrete options, rather than using averaging to integrate continuous judgments. Wisdom-of-crowds effects are typically studied in the latter type of task (Galton, 1907; Surowiecki, 2004), although it has been shown theoretically that the performance of majority and plurality rules often compares to that of a computationally more demanding averaging rule (Hastie & Kameda, 2005). In further work, tasks that involve choice between accepting or rejecting a given option can be modeled using signal detection theory, following for example, Sorkin et al. (1998). However, as discussed earlier, the effects of averaging over different task difficulties are likely to remain even after accounting for individual differences in detection sensitivities of group members on a particular task.

Finally, note that we did not assume any selective sampling of group members on the basis of, for example, expertise (Budescu & Chen, 2015; Goldstein et al., 2014; Mannes et al., 2014). A smaller group that would produce more accurate decisions in our model can simply be selected randomly out of a larger group of experts. More generally, our results suggest that even though modern technologies enable easier communication in large groups, the resulting decisions may be (sometimes drastically) less accurate than those that would have been made in moderately sized groups with the same average individual accuracy. Institutional designers in government and industry can consider these results when determining the best committee size for the range of tasks their experts will have to face.

\begin{thebibliography}{99}


\end{thebibliography}


Received July 20, 2015
Revision received December 23, 2015
Accepted April 3, 2016