Can Small Crowds Be Wise? Moderate-Sized Groups Can Outperform Large Groups and Individuals Under Some Task Conditions

Mirta Galesic
Daniel Barkoczi
Konstantinos Katsikopoulos

SFI WORKING PAPER: 2015-12-051

SANTA FE INSTITUTE

Can small crowds be wise?
Moderate-sized groups can outperform large groups and individuals under some task conditions

Mirta Galesic[12*], Daniel Barkoczi[2], & Konstantinos Katsikopoulos[2]

[1]Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501, USA

[2]Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, 14195, Berlin, Germany

[*]Corresponding author: galesic@santafe.edu

Keywords: wisdom of crowds; majority rule; Condorcet Jury Theorem; group decision making

Abstract

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, and that technological advances make group meetings easier than ever before, 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 is above chance, but some are surprisingly difficult, with most group members being wrong. Examples of the latter are elections with unexpected 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 in these circumstances moderate-sized groups can achieve higher average accuracy across all tasks than larger groups or individuals. This happens because an increase in group size can lead to a decrease in group accuracy for difficult tasks which is larger than the corresponding increase in accuracy for easy tasks. We derive this non-monotonic relationship between group size and accuracy from Condorcet Jury Theorem and use simulations and further analyses to show that it holds 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 finite populations without replacement. We further show that situations favoring moderate-sized groups occur in a variety of real-life domains including political, medical, and financial decisions, and general knowledge tests. We discuss implications for the design of decision-making bodies at all levels of policy.

Introduction

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 (U.K. Department for Communities and Local Government, 2008; Electoral Council of Australia & New Zealand, 2013), 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 (Lyberk & Morris, 2004). Similarly, individuals considering a variety 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 & Rukton, 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 meeting and communication 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). We complement these explanations with an argument for the superiority of moderate group sizes based solely on group decision accuracy.

In many real-life situations, expert groups encounter mostly easy tasks on which average individual accuracy is above chance, and some surprisingly difficult tasks where most members guess wrongly (see next section for examples). Here we show that, when it is *not known whether the next task will be easy or difficult*, average decision accuracy peaks when voting is done by moderately sized groups. This does not occur because of selective sampling of group members based on expertise (Budescu & Chen, 2014; Goldstein, McAfee, & Suri, 2014; Mannes, Soll, & Larrick, 2014) but solely because the accuracy of groups deciding by simple majority or plurality rules increases with their size for relatively easy tasks but decreases for tasks for which most individuals make the wrong prediction.

Tasks with Surprising Outcomes

Tasks with unexpected outcomes that are difficult to predict can be found in many domains, including political and economic forecasts, medical diagnoses, and general knowledge tests. For example, 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 U.K. 2015 general election, where all but one polling company erroneously predicted that Tories would not win a majority of seats in the Parliament (Bialik, 2015). Similarly, majority of forecasters in the U.S. 2000 presidential

elections predicted Gore's victory over Bush in Florida (Graefe, 2014; Whitson, 2001). As illustrated in the last section, experts such as political forecasters, medical doctors, financial experts, or trivia quiz participants, who do not know whether the next task will be easy or difficult, will often do best to decide by majority in moderately sized groups rather than in large groups or individually.

Consider the knowledge question "Which city is farther north, New York or Rome?", which most people answer incorrectly. Temperature, the cue that is valid for most other comparisons of city latitudes, points to the wrong answer for this pair of cities (Gigerenzer, Hoffrage, & Kleibölting, 1991). In cases such as these, majority of individuals can be wrong, resulting in average individual accuracy below 50% on those particular tasks. This can happen because these 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.

However, 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 has changed between the moment of prediction and the moment when the outcome was observed. For instance, most diseases might be accurately diagnosed based on 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 years because of

unobservable underlying complexities affecting the financial markets. In what follows, we will call tasks with surprising outcomes that most people predict incorrectly "difficult", and those that most people predict correctly "easy".

Group Size and Accuracy over a Range of Task Difficulties

Most committees will face a variety of task difficulties in the course of their existence, ranging from very easy to quite difficult. However, most past studies of group decision accuracy have assumed that groups always encounter tasks of the *same* and *known* difficulty. 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). However, in most real-life situations 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 quite easy tasks and occasionally some surprisingly difficult tasks. A novice group might find most tasks very difficult and some quite easy. In addition, in some domains predictions are inherently easier than in others. Not knowing exactly what task difficulties a group will face, can we 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 real-life situations moderate-sized groups will achieve higher accuracy than 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, and where it is eventually possible to determine whether the group decision was correct or not (see real-world examples in the last section). Note that the voting stage may or may not be preceded by a group discussion where members determine common ground for understanding the problem, share some or all of the information they possess individually, make various quantitative 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, where individual votes are transformed into a group vote for one of two or more possible courses of action.

For simplicity, we first analyze situations with tasks involving two options between which groups choose by simple majority rule, considering only two task difficulties, assuming that individual group members vote independently, and assuming that they are selected from a very large population with replacement. Afterwards, 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. In the last section, we provide several real-world examples from different task domains where smaller groups can perform better than larger ones.

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, we can use the Condorcet Jury Theorem (CJT), which can be represented as

$$P_n = \sum_{i=m}^{n} \binom{n}{i} \overline{p}^i (1-\overline{p})^{n-i} \qquad [1]$$

where P_n is group accuracy at group size n, m is size of simple majority, and $\overline{p}$ is average individual accuracy. Without loss of generality, n is assumed to be always odd. Individual group members can have heterogeneous skills.[1] 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 (Sorkin, West, & Robinson, 1998).

To study average group accuracy over two or more tasks, we first assume that groups encounter two task difficulties: With probability e they encounter easy (denoted E) tasks, for which average individual accuracy $\overline{p}_{\mathrm{E}} > 0.5$; and with probability 1 - e they encounter surprising or difficult (denoted D) tasks, for which average individual accuracy $\overline{p}_{\mathrm{D}} < 0.5$. Figure 1 shows how average group accuracy $\overline{P}$ 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 (Eq. 1), for easy tasks group accuracy P_{E} is larger than $\overline{p}_{\mathrm{E}}$ and increases monotonically to 1 as groups get larger (red dashed lines in all panels of Figure 1). For difficult tasks, $P_{\mathrm{D}} < \overline{p}_{\mathrm{D}}$ and decreases monotonically to 0 with increase in group size (blue

[1] 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 et al, 1983). Deviations occur only in exceptional cases, for instance when some individuals consistently have accuracy 0 or 1 or when average accuracy is close to 0.5 and groups are very small. With increase in n, group accuracy P monotonically increases to 1 for tasks with average individual accuracies $\overline{p} > 0.5$ and monotonically decreases to 0 for tasks with $\overline{p} < 0.5$. When average of individual accuracies $\overline{p} = 0.5$, P will converge to a value between 0.39 and 0.61 (Owen, Grofman, & Feld, 1989). In other words, CJT predictions generalize to a large range of asymmetrical distributions of individual skills (Grofman et al., 1983).

dotted lines). The average group accuracy $\overline{P}$ (full black 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:

$$\bar{P}_n = e\ P_{\mathrm{E},n} + (1-e)P_{\mathrm{D},n} \qquad [2]$$

where $\bar{P}_n$ is the average accuracy of a group of size *n*, *e* is the proportion of easy tasks, and $P_{\mathrm{E},n}$ ($P_{\mathrm{D},n}$) is the accuracy of a group of size *n* on easy (difficult) tasks derived by the CJT.

As Figure 1 illustrates, changes in $\overline{P}$ 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 encountered by a group of experts who are skilled in solving particular tasks and, even when surprised, don't 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, as specified below.

More formally, we define a "friendly" environment as one in which $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} > 1$, a "neutral" environment as one in which $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} = 1$, and an "unfriendly" environment as one in which $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} < 1$. 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, $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} > 1$ is equivalent to $\overline{p}_{\mathrm{E}} - 0.5 > 0.5 - \overline{p}_{\mathrm{D}}$, which means that in friendly environments, the accuracy in easy tasks is above chance more than the accuracy in difficult tasks is below chance.

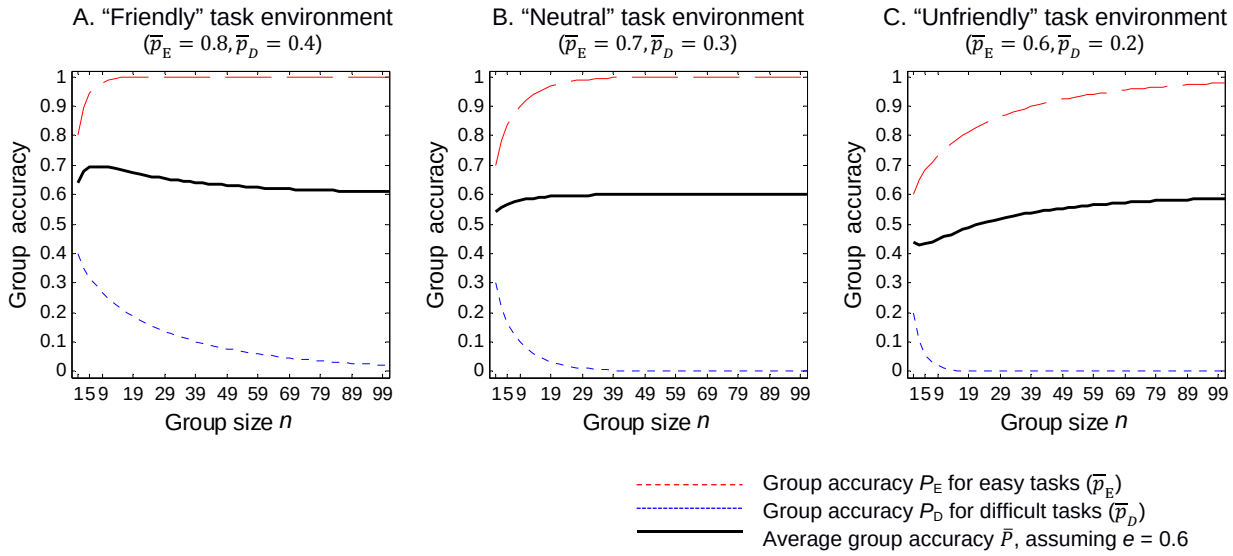

Figure 1. Average group accuracy can peak at moderate group sizes. Illustration of changes in group accuracy as a function of group size *n* and different combinations of task difficulties, assuming proportion of easy tasks *e* = 0.6. Note that as *n* increases, average group accuracy $\overline{P}$ converges to *e*. (**A**) In a "friendly" task environment ($\overline{p}_E + \overline{p}_D > 1$), $\overline{P}$ increases until *n* = 7, then decreases toward *e.* (**B**) In a "neutral" task environment ($\overline{p}_E + \overline{p}_D = 1$), $\overline{P}$ increases monotonically with *n* until it reaches *e*. (**C**) In an "unfriendly" task environment ($\overline{p}_E + \overline{p}_D < 1$), $\overline{P}$ decreases until *n* = 3, then increases toward *e* = 0.6 even though $\overline{P}_{n=1} < 0.5$.

In all environments, $\overline{P}_n$ will start from $\overline{P}_{n=1} = e\overline{p}_E + (1-e)\overline{p}_D$, which is the average individual accuracy across easy and difficult tasks, and with increase in *n* 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 $\overline{P}$ converges to $e \times 1 + (1-e) \times 0 = e$.

In between these two extremes, $\overline{P}_{n=1}$ and *e*, $\overline{P}$ can be a monotonically increasing, monotonically decreasing, U-shaped, or inverted-U-shaped function of *n*. Which of these shapes obtains is completely determined by two factors defined precisely above: the type of environment (friendly, neutral, or unfriendly) and the value of the starting point $\overline{P}_{n=1}$.

Note that because $\overline{P}_{n=1} = e\overline{p}_{\mathrm{E}} + (1-e)\overline{p}_{\mathrm{D}}$, the condition $\overline{P}_{n=1}$>0.5 can be equivalently expressed as $e > (0.5 - \overline{p}_{\mathrm{D}})/(\overline{p}_{\mathrm{E}} - \overline{p}_{\mathrm{D}})$.

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

$$\Delta\bar{P}_n \begin{cases} > 0 & \text{if } \Delta P_{\mathrm{E},n} > \frac{1-e}{e}\Delta P_{\mathrm{D},n} \\ < 0 & \text{if } \Delta P_{\mathrm{E},n} < \frac{1-e}{e}\Delta P_{\mathrm{D},n} \\ = 0 & \text{if } \Delta P_{\mathrm{E},n} = \frac{1-e}{e}\Delta P_{\mathrm{D},n} \end{cases} \qquad [3]$$

where $\Delta\bar{P}_n = \bar{P}_{n+2} - \bar{P}_n$ is change in average group accuracy across all tasks, and $\Delta P_{\mathrm{E}} = P_{\mathrm{E},n+2} - P_{\mathrm{E},n}$ and $\Delta P_{\mathrm{D}} = P_{\mathrm{D},n+2} - P_{\mathrm{D},n}$ represent change in average group accuracy across easy and difficult tasks, respectively. In words, average group accuracy $\overline{P}$ will increase with group size if the rate of change in accuracy on easy tasks $\Delta P_{\mathrm{E},n}$ is higher than the rate of change in accuracy on difficult tasks $\Delta P_{\mathrm{D},n}$, weighted by the relative prevalence of difficult tasks $\frac{1-e}{e}$. 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}, $\overline{P}$ will increase and otherwise decrease. It will reach its peak when the gains and weighted losses cancel each other.

To illustrate Eq. 3, consider a friendly environment, in which $\overline{p}_{\mathrm{E}} - 0.5 > 0.5 - \overline{p}_{\mathrm{D}}$. Here, the rate of change in accuracy on easy tasks is initially higher than the rate of change on difficult tasks, as follows from Eq. 1 when $\overline{p}_{\mathrm{E}}$ is closer to 1 than $\overline{p}_{\mathrm{D}}$ is to 0. When in addition easy tasks are encountered more often, that is when e > 0.5 and thus $\frac{1-e}{e} < 1$, the difference in rates of change is further magnified and $\overline{P}$ will definitely increase with group size. On the other hand, when easy tasks are encountered less often, that is when e < 0.5 and thus $\frac{1-e}{e} > 1$, $\overline{P}$ may not increase with group size even if the environment is friendly. Importantly, even with e > 0.5, the initially increasing trend in $\overline{P}$ may be reversed as n

continues to increase because P_E will reach its limiting value, 1, while P_D will still be decreasing towards zero, driving $\overline{P}$ down. This is what happens in Figure 1A. In this friendly task environment, an increase in *n* initially leads to an increase in $\overline{P}$, here peaking at 0.7 for $n = 7$ before decreasing to *e*.

Similar analysis can be applied to other environments. The component of $\overline{P}$, eP_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 $\overline{P}$ increases monotonically with *n* until it reaches *e*. Finally, Figure 1C shows a particularly interesting case that occurs in unfriendly environments. Here, a downward peak occurs, with $\overline{P}$ initially decreasing and then slowly increasing toward *e*. Note that in this case $\overline{P}$ 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 ($\overline{P}_{n=1} = 0.44$).

Solving Eq. 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). To examine how changes in small *n* relate to group accuracy for different combinations of task difficulties, we calculated $\bar{P}$ using Eq. 2 across a range of group sizes, for all combinations of easy ($0.6 \leq \overline{p}_E \leq 0.9$) and difficult ($0.1 \leq \overline{p}_D \leq 0.4$) tasks, separately for different proportions of easy tasks $0.1 \leq e \leq 0.9$, in increments of 0.1. Results presented in Figures 2 and S1 show that non-monotonic changes in $\overline{P}$, such as those shown in Figure 1, occur in more than half of all possible combinations of task difficulties.

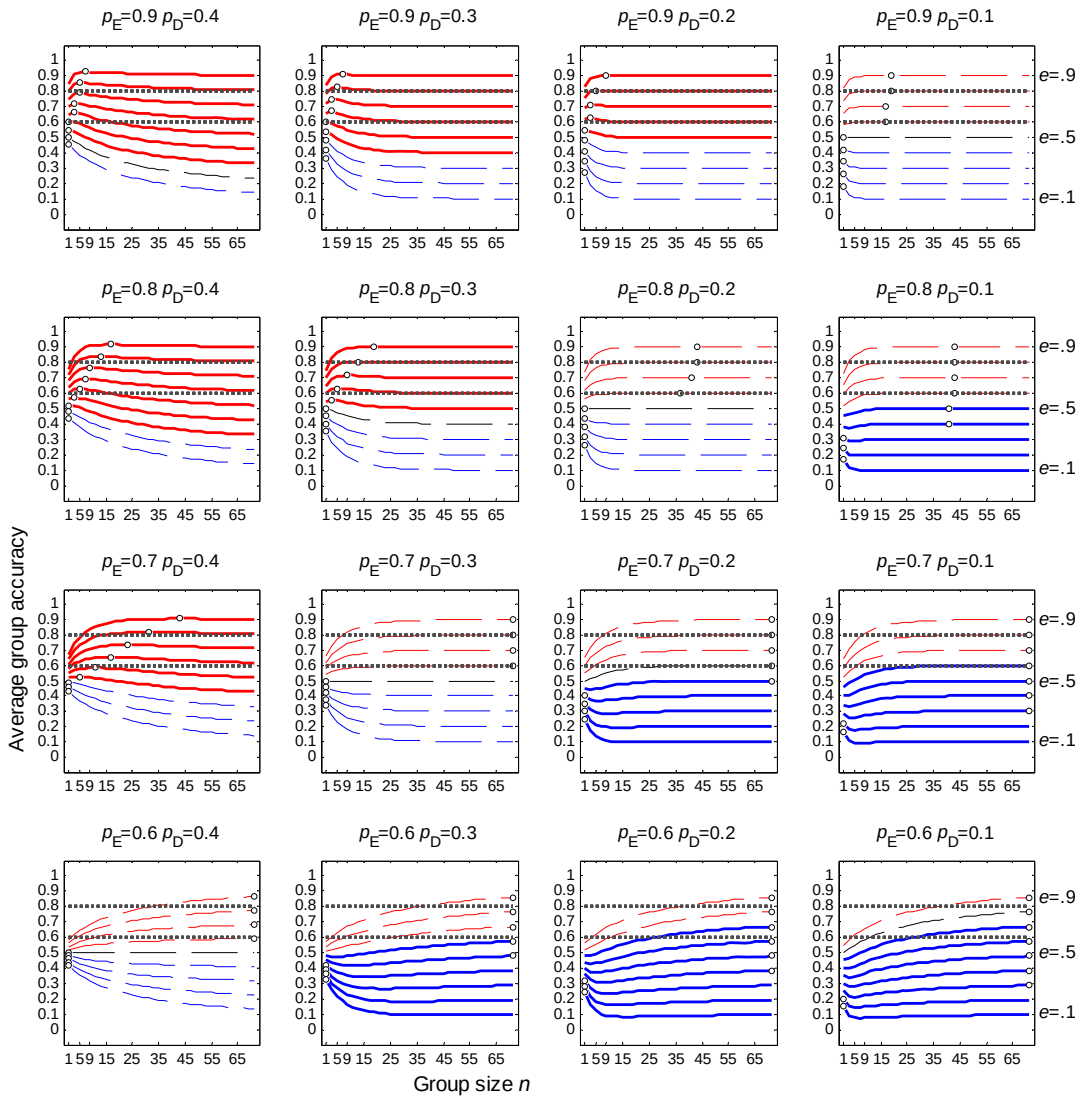

Figure 2. Average group accuracy depends on combination of task difficulty and proportion of easy tasks. Each panel shows changes in average group accuracy $\bar{P}$ as a function of group size *n*, different combinations of easy ($0.6 \leq \overline{p}_{\mathrm{E}} \leq 0.9$) and difficult ($0.1 \leq \overline{p}_{\mathrm{D}} \leq 0.4$) tasks, and different proportions of easy tasks ($0.1 \leq e \leq 0.9$). Red lines represent cases in which average individual accuracy across tasks $\overline{P}_{n=1} > 0.5$, blue lines are for $\overline{P}_{n=1} < 0.5$, and black lines for $\overline{P}_{n=1} = 0.5$, where $\overline{P}_{n=1}=e\overline{p}_{\mathrm{E}} + (1-e)\overline{p}_{\mathrm{D}}$. Circles show maximum value of $\bar{P}$ for each case. Dashed lines denote cases where $\bar{P}$ changes monotonically with *n* until it reaches *e*, while solid lines denote cases where $\bar{P}$ changes nonmonotonically, that is, reaches an upward or a downward peak at moderate group size *n* before reaching *e*. 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 (see main text for details). Gray dotted lines denote region in which $0.6 \leq \overline{P}_{n=1} \leq 0.8$, as is commonly observed in real-world policy tasks.

Moderate group sizes have advantage over larger groups or single individuals in all friendly environments ($\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} > 1$, subplots above the diagonal) whenever proportion of easy tasks is $e > (0.5 - \overline{p}_{\mathrm{D}})/(\overline{p}_{\mathrm{E}} - \overline{p}_{\mathrm{D}})$, that is, when average individual accuracy across tasks is larger than chance ($\overline{P}_{n=1} > 0.5$). In addition, moderate group sizes are as good as larger group sizes in neutral ($\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} = 1$, subplots on the diagonal) and unfriendly environments ($\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} < 1$, subplots below the diagonal) whenever easy tasks are very easy ($\overline{p}_{\mathrm{D}}$>= 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.

Figure 2 further shows how average group accuracy $\overline{\mathrm{P}}$ changes with increase in group size n with friendliness of the task environment (as reflected in the sum of average individual accuracy on easy and difficult tasks $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}}$), and average individual accuracy across tasks ($\overline{P}_{n=1}$). Specifically, (i) when $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} > 1$ and $\overline{P}_{n=1} > 0.5$, $\overline{P}$ will reach an upward peak at moderate n; (ii) when $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} \geq 1$ and $\overline{P}_{n=1} < 0.5$, $\overline{P}$ will decrease monotonically with n towards e; (iii) when $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} \leq 1$ and $\overline{P}_{n=1} > 0.5$, $\overline{P}$ will increase monotonically with n towards e; and (iv) when $\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} < 1$ and $\overline{P}_{n=1} < 0.5$, $\overline{P}$ will reach a downward peak at moderate n.

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 with surprising outcomes. In the following sections, we add further realistic assumptions and examine real-world situations.

Before proceeding, note that in this paper we formally test the verbal conjecture made by Grofman et al. (1984) that nonmonotonicity in average proportion correct across different sample sizes can be expected in multi-item tasks with hard and easy items. We provide several novel results that were not anticipated or described by Grofman et al. First, we define exact conditions when average accuracy over several tasks will increase with group size, when it will decrease, and when it will achieve a peak, rather than stating only verbally that these changes are expected to be non-monotonic in some circumstances (Eq. 3 above and related discussion). Second, we show that moderately-sized groups can be preferable to larger groups even without 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 (Figure 2). Fourth, we delineate conditions for non-monotonic trends in group accuracy with downward peaks, that is when moderate group sizes are less accurate than both single individuals and large groups (see above). Fifth, in what follows, 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 moderate-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 $\overline{p}_{\mathrm{E}}$ and $\overline{p}_{D}$ for all easy and difficult tasks, respectively (although on each task individuals could have heterogeneous skills). In real life, groups will face tasks of a wide range of difficulties. Average group accuracy across many different tasks can be calculated by an extension of Eq. 2:

$$\overline{P}_{n} = \frac{1}{T}\sum_{t=1}^{T} P_{t,n} \qquad [4]$$

where T is the number of tasks, and $P_{t,n}$ is group accuracy on a given task t at group size n, calculated using Eq. 1.

More generally, instead of assuming that task difficulties $\overline{p}_{\mathrm{E}}$ and $\overline{p}_{D}$ are the same for all easy and difficult tasks that a group encounters, we can model them as random draws from beta distributions with parameters $\alpha_E = \overline{p}_{\mathrm{E}}k$ and $\beta_E = (1-\overline{p}_{\mathrm{E}})k$ for easy tasks, and parameters $\alpha_D = \overline{p}_{\mathrm{D}}k$ and $\beta_D = (1-\overline{p}_{\mathrm{D}})k$ for difficult tasks, where k is a constant that determines the size of the variance of task difficulties. Then, easy tasks have mean difficulty $\alpha_E/(\alpha_E+\beta_E) = \overline{p}_{\mathrm{E}}$ and variance $\alpha_E\beta_E/(\alpha_E+\beta_E)^2(\alpha_E+\beta_E+1) = \overline{p}_{\mathrm{E}}(1-\overline{p}_{\mathrm{E}})/(k+1)$. Similarly, difficult tasks can be modeled as having a mean $\alpha_D/(\alpha_D+\beta_D) = \overline{p}_{\mathrm{D}}$ and variance $\alpha_D\beta_D/(\alpha_D+\beta_D)^2(\alpha_D+\beta_D+1) = \overline{p}_{D}(1-\overline{p}_{D})/(k+1)$. As k increases, the variance decreases.

To check how assuming a range of two task difficulties affects average group accuracy, we replicated the simulations above for different average task difficulties as before, assuming different levels of variance of task difficulties: small (k=100), moderate (k=50), and large (k=10). Figures S2A-S4B show that the results described above hold even

when distributions of task difficulties have large variances, though the results become more noisy.

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 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, ..., p_k that an average individual chooses each of the k options. Once group accuracies $P_{t,n}$ are calculated in this way for different tasks t and group sizes n, Eq. 4 can be used to calculate average group accuracy. It is then easy to show that nonmonotonic changes in average group accuracy $\bar{P}$ can occur in these situations, as well.

Effect of Correlated Votes

So far we have assumed that group members are independent in a 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., Bromell & 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 (Kao & Couzin, 2014; Boland, Proschan, & Tong, 1989; Spiekermann & Goodin, 2012).

These effects of correlated votes can be parsimoniously explained within the present framework. Whenever the leader or the common information 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, given stochastic accuracy of the leader or the common cue, the overall group accuracy can be represented as an average of its performance on easy and difficult tasks. Accordingly, single peak functions of the kind presented above have been observed for groups with correlated votes (see references above) but to our knowledge the simple explanation in terms of a mixture of easy and difficult tasks has not been proposed before.

More formally, following an opinion leader (who does not vote but influences some group members to decide in a certain way) or voting based on 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,

$$\overline{P}_n = lP_n[p(1-r)+r] + (1-l)P_n[p(1-r)] \qquad [5]$$

where $\overline{P}_n$ is average accuracy of a group of size n, l is probability that an opinion leader is accurate on a certain task, P_n is group accuracy at group size n given individual accuracy specified within the square brackets, p is initial individual accuracy of group members, and r is the proportion of group members who are following the opinion leader. The higher r, the higher the correlation among group members, and in some cases the two values are

identical (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. A condition similar to Eq. 3 must be satisfied for $\bar{P}$ to increase with group size n:

$$\overline{P}_{n+2}[p(1-r)+r] - \overline{P}_{n}[p(1-r)+r] > \frac{1-l}{l}\left[\overline{P}_{n+2}[p(1-r)] - \overline{P}_{n}[p(1-r)]\right] \qquad [6]$$

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 the results presented above, 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, Eqs. 2 and 5 can be combined to account for both correlated votes and task difficulty:

$$\overline{P}_{n} = e\left[l_{\mathrm{E}}P_{\mathrm{E},n}\left[\overline{p}_{\mathrm{E}}\,(1-r)+r\right] + (1-l_{\mathrm{E}})P_{\mathrm{E},n}\left[\overline{p}_{\mathrm{E}}\,(1-r)\right]\right] +$$

$$+(1-e)\left[l_{\mathrm{D}}P_{\mathrm{D},n}\left[\overline{p}_{D}\,(1-r)+r\right] + (1-l_{\mathrm{D}})P_{\mathrm{D},n}\left[\overline{p}_{\mathrm{D}}\,(1-r)\right]\right] \qquad [7]$$

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

We repeat the analyses above (presented in Figures 2 and S1B) while increasing the assumed proportion of voters *r* who follow the leader from 0 to 1 in steps of 0.1.[2] With

[2] In these simulations, we assumed that the leader has the same skill as the average group member ($l_{\mathrm{E}} = \overline{p}_{\mathrm{E}}$ and $l_{\mathrm{D}} = \overline{p}_{\mathrm{D}}$). The results still hold if we assume 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.

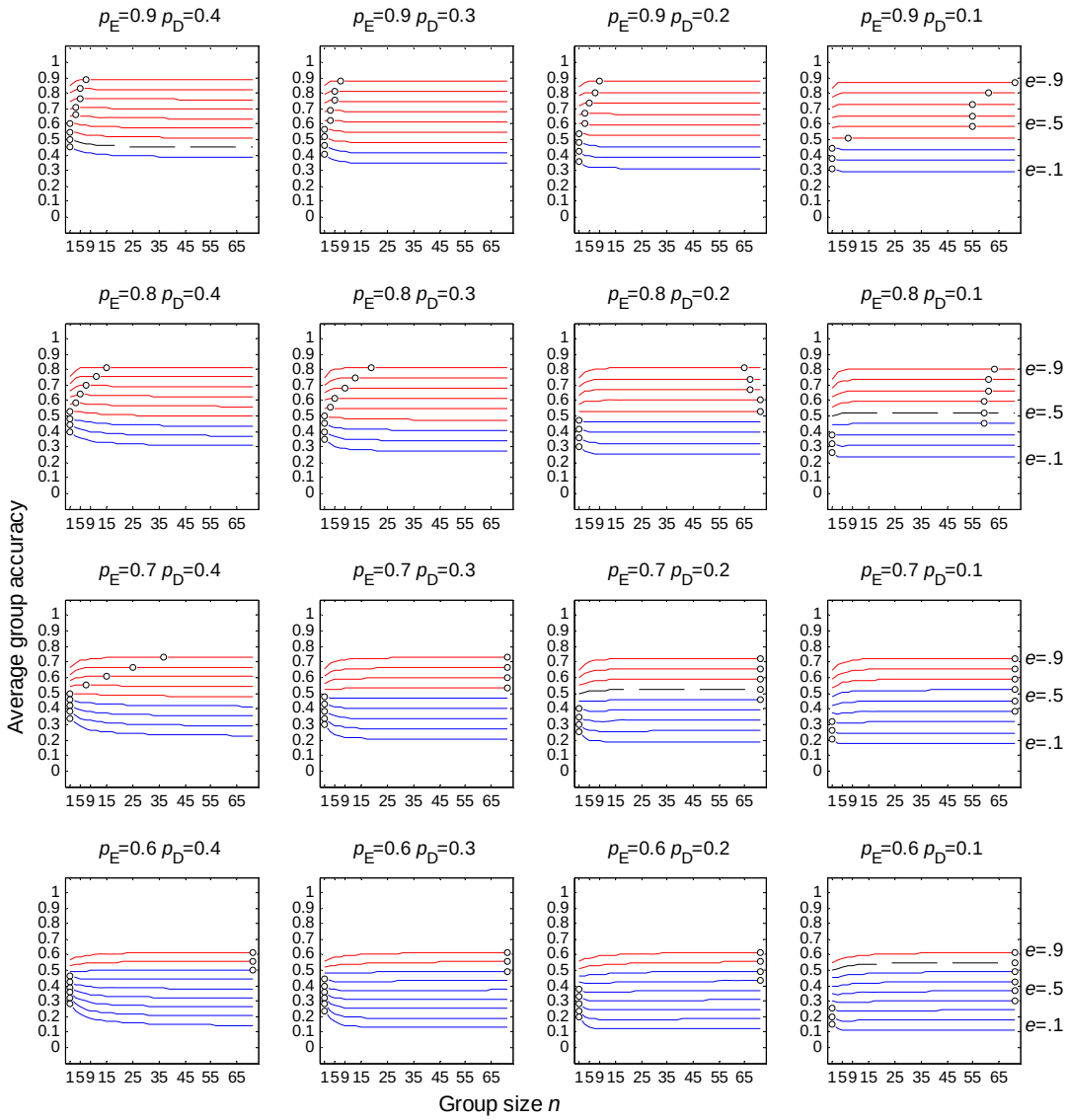

Figure 3. Group accuracy after repeating the analysis in Figure 2 for correlation levels *r* ranging from 0 to 1 in steps of 0.1 and averaging over them. Leader accuracy l is assumed to be equal to the average individual accuracy $\overline{p}$ (see text for more details).

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 (Hogarth, 1978). 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 average the results over the whole range of values of *r*. The results, shown in Figures 3 and S5, 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, Eq. 1 could be rewritten using cumulative hypergeometric rather than cumulative binomial distribution:

$$P_n = \sum_{i=m}^{n} \frac{\binom{I}{i}\binom{N-I}{n-i}}{\binom{N}{n}} \qquad [8]$$

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 from which the smaller group of experts is randomly selected), $I=N*i/n$, and I/N equals $\overline{p}$, 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 above to enable comparison of our results with previous studies, but to check whether any conclusions presented in the paper would be different when using hypergeometric distribution, we re-run all the analyses using that statistical model. Figures S6 and S7 show the results assuming that members of smaller groups were randomly selected from finite populations without replacement: Figure S6 assumes a population of

N=71, and Figure S7 assumes *N*=31. Both analyses suggest that the finding that group accuracy was often maximized for moderate-sized groups is, if anything, more pronounced with these more realistic assumptions.

Real-World Illustrations

What is the best committee size in real-world environments? To answer this question, we need 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, we could expect that they are on average more accurate than chance. In addition, we 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. To illustrate, 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 ($\overline{P}_{n=1} = 0.66$). In easy years, average individual accuracy $\overline{p}_{\mathrm{E}}$ was 0.88, and in two difficult years (Bush vs. Gore, 2000; and Bush vs. Kerry, 2004), average individual accuracy $\overline{p}_{\mathrm{D}}$ 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 ($\overline{P}_{n=1} = 0.70$). For diseases that were easy to diagnose average individual accuracy $\overline{p}_{\mathrm{E}}$ was 0.81, and for difficult ones, including Lyme disease, pyrogenic spinal infections, and abdominal aortic aneurysm, $\overline{p}_{\mathrm{D}}$

was 0.41 (Schiff et al., 2009; 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 ($\overline{P}_{n=1} = 0.71$). Two of the domains were relatively easy to predict ($\overline{p}_{\mathrm{E}} = 0.86$), while economic growth was somewhat difficult ($\overline{p}_{\mathrm{D}} = 0.43$; Hilsenrath & Peterson, 2013; 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 ($\overline{P}_{n=1} = 0.76$), 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 $\overline{p}_{\mathrm{E}}$ was 0.86, and on the difficult tasks $\overline{p}_{\mathrm{D}}$ was 0.38. In sum, in all of these examples task environments were friendly ($\overline{p}_{\mathrm{E}} + \overline{p}_{\mathrm{D}} > 1$), and each expert had above chance accuracy on an average task ($\overline{P}_{n=1} > 0.5$). These conditions satisfy the conditions outlined above for the situations when groups of moderate sizes are likely to reach the highest accuracy.

If we assume that a policy maker 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 her 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 use Eq. 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 above (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 paper two of those seven seats are empty (Federal Reserve, December 2015). Finally, for answering general knowledge items correctly, the best group size for participants of Juslin's (1994) study is $n = 15$.

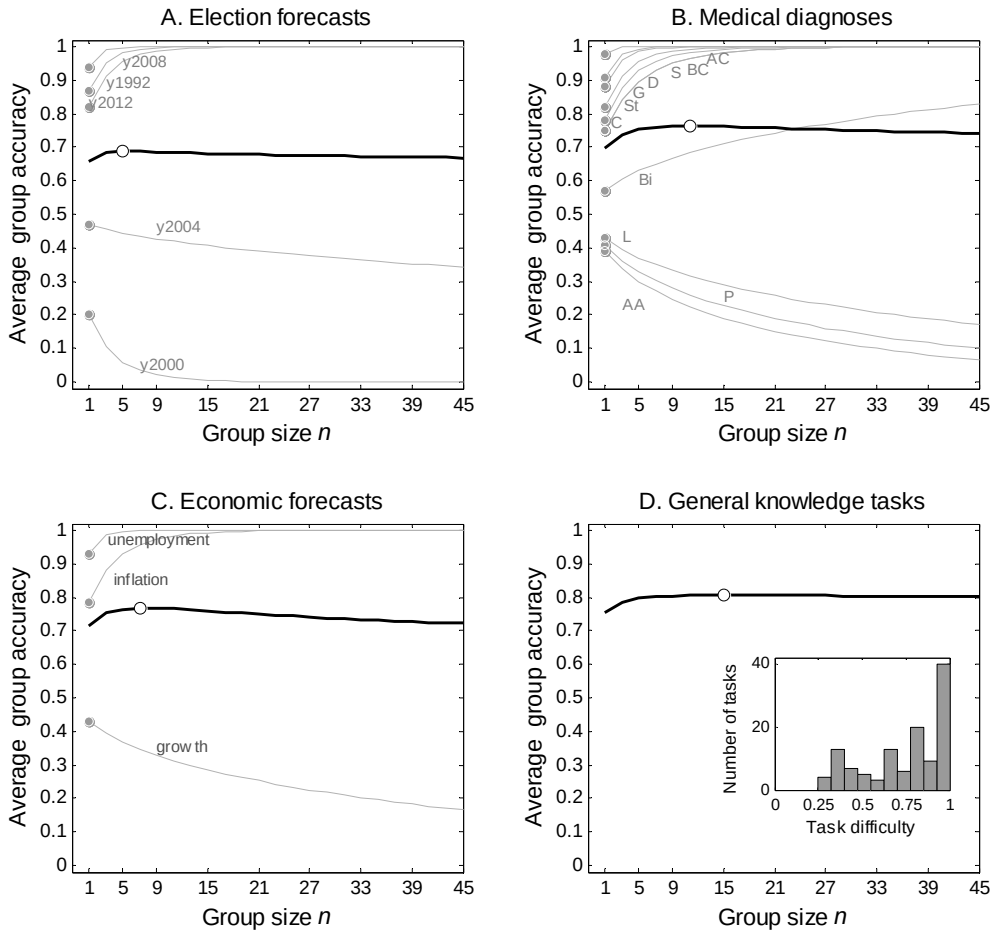

Figure 4. Real-world environments are often friendly and group accuracy peaks at moderate group sizes. Gray dots in (A-C): Average individual accuracies for particular tasks (five election forecasts in A, diagnoses for 11 diseases in B, forecasts for three economic trends in C). **Inset in (D):** Histogram of average individual accuracies for 120 knowledge tasks. **Gray lines:** Group accuracy for different group sizes, for each of the different tasks faced by (**A**) experts predicting U.S. political elections in years 1992, 2000, 2004, 2008, and 2012 (Graefe, 2014), (**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), (**C**) U.S. Federal Reserve Bank officials giving economic forecasts about future economic trends in unemployment (unempl), inflation, and economic growth (Hilsenrath & Peterson, 2013), and (**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 (**D**) each gray line depicts several tasks and frequency of different tasks at each level of task difficulty ($\overline{p}$) 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:** average group accuracy across different tasks. In all four examples, average group accuracy peaks at moderate group sizes (as indicated by circles): in **A** at n = 5; in **B** at n = 11; in **C** at n = 7; and in **D** at n = 15.

Discussion

Our results suggest that the highest accuracy across a diverse set of tasks involving choice between two or more courses of action may 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 we assume 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.

While group size that achieves highest accuracy depends on the individual accuracy on easy and difficult tasks and the proportion of easy tasks, we show that conditions favoring relatively small committees may 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. While real-world group sizes are influenced by many factors other than accuracy, our results show that groups that are smaller because of organizational or communication constrains 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. Unless it is somehow cheaper to support larger rather than smaller groups of otherwise comparable individuals, it will always be more efficient to have a moderate-sized rather than a larger committee.

Note that we modeled tasks in which groups use simple majority or plurality rules to choose between discrete options, rather than using averaging to predict a quantitative property. 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 which involve choice between accepting or rejecting a given option can be modeled using signal detection theory, following e.g. Sorkin et al (1998). However, as discussed above, 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 do not assume any selective sampling of group members, for example, based on expertise (Budescu & Chen, 2014; Goldstein, McAfee, & Suri, 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.

Acknowledgments

Matlab scripts for all calculations are available from the authors. We thank the Max Planck Institute for Human Development and the Santa Fe Institute for their support; David Budescu, Reid Hastie, John Miller, Shenghua Luan, Henrik Olsson, and Scott Page for helpful comments on an earlier version; and Anita Todd for editing the manuscript.

References

Anderson, M. (2014). Local Consumer Review Survey 2014. Retrieved from: http://www.brightlocal.com/2014/07/01/local-consumer-review-survey-2014/

Bialik, C. (2015, May 13). SurveyMonkey was the other winner of the U.K. election. *FiveThirtyEight*. Retrieved from http://fivethirtyeight.com/features/surveymonkey-was-the-other-winner-of-the-u-k-election/.

Boland, P.J., Proschan, F., Tongm Y.L. (1989). Modelling dependence in simple and indirect majority systems. *Journal of Applied Probability, 26,* 81-88.

Broomell, S. B., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika,74*, 531–553.

Budescu, D.V., Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science.* http://dx.doi.org/10.1287/mnsc.2014.1909

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix [Essay on the application of analysis to the probability of majority decisions].*Paris: Imprimerie Royale.

Dejure.org (2013). Betriebsverfassungsgesetz: Zahl der Betriebsratsmiglieder. Retrieved from http://dejure.org/gesetze/BetrVG/9.html.

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*, 681-735.

Electoral Council of Australia & New Zealand (2013). Electoral systems of Australia's Parliaments and local government. Retrieved from http://www.eca.gov.au/systems/files/1- electoral-systems.pdf.

European Parliament (2014). List of committees. Retrieved from http://www.europarl.europa.eu/committees/en/home.html.

Federal Reserve (2014). Board of Governors of the Federal Reserve System. Retrieved from http://www.federalreserve.gov/aboutthefed/default.htm.

Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. *Psychological Science, 23*, 1515-1523.

Galton, F. (1907) Vox populi. *Nature, 75,* 450-451.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 254-267.

Goldstein, D.G., McAfee, P., Suri, S. (2014). The wisdom of smaller, smarter crowds. Proceedings of the 15th ACM Conference on Economics and Computation, 471-488.

Graefe, A. (2014). Accuracy of vote expectation surveys in forecasting elections. *Public Opinion Quarterly, 78*, 204-232.

Grofman, B., Owen, G., & Feld, S.L. (1983). Thirteen theorems in search of the truth. *Theory and Decision, 15*, 261-278.

Grofman, B., Feld, S.L., & Owen, G. (1984). Group size and the performance of a composite group majority: Statistical truths and empirical results. *Organizational Behavior and Human Performance, 33,* 350-359.

Haas, L. K. (2014). List of standing committees and select committee and their subcommittees of the House of Representatives of the United States. U.S. House of Representatives: Office of the Clerk. Retrieved from http://clerk.house.gov/committee_info/scsoal.pdf.

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

Hilsenrath, J. & Peterson, K. (2013). Federal Reserve 'doves' beat 'hawks' in economic prognosticating. *The Wall Street Journal*, July 29, 2013. Retrieved from http://online.wsj.com/news/articles/SB10001424127887324144304578624033540135700.

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

Juslin, P. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104*, 344-366

Kao, A.B., Couzin, I.D. (2014). Decision accuracy in complex environments is often maximized by small group sizes, *Proceedings of Royal Society B*, *281*, 2013305.

Krause, J. & Ruxton, G.D. (2002). *Living in Groups*. Oxford, UK: Oxford Univ. Press.

Leib, E. J. (2008). A comparison of criminal jury decision rules in democratic countries. *Ohio State Journal of Criminal Law, 5*, 629-644.

LiquidFeedback; http://liquidfeedback.org (2014). Date of access: 2014.09.03.

List, C. & Goodin, R.E. (2001). Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9, 277–306.

Lybek, T. & Morris, J. (2004). Central bank governance: A survey of boards and management. International Monetary Fund, WP/04/226.

Mannes, A.E., Soll, J.B., Larrick, R.P (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*, 276-299.

Owen, G., Grofman, B., & Feld, S.L. (1989). Proving a distribution-free generalization of the Condorcet Jury Theorem. *Mathematical Social Sciences, 17*, 1-16.

Parliament of Australia (2014). House of Representatives – Committees. Retrieved from http://www.aph.gov.au/parliamentary_business/committees/house_of_representatives_committees?url=comm_list.htm.

Schiff, G.D. et al. (2009). Diagnostic error in medicine: Analysis of 583 physician-reported errors. *Archives of Internal Medicine, 169,* 1881-1887.

Sorkin, R.D., West, R., & Robinson, D.E. (1998). Group performance depends on the majority rule. *Psychological Science, 9,* 456-463.

Spiekermann, K. & Goodin, R.E. (2012). Courts of many minds. *British Journal of Political Science, 42,* 555-571.

Surowiecki, J. (2004). *The wisdom of crowds.* New York: Doubleday.

Tideman, T. N., & Plassmann, F. (2013). Developing the empirical side of computational social choice. *Annals of Mathematics and Artificial Intelligence, 68*, 31-64.

UK Department for Communities and Local Government (2008). Guidance on community governance reviews. Retrieved from http://www.nalc.gov.uk/Document/Download.aspx?uid=f2136ed5-fe2f-4cfa-8058-c3cbd404c987.

Whitson, J.R. (2001). The 2000 Electoral College predictions scoreboard. Retrieved from http://presidentelect.org/art_2000score.html.

Supplemental Materials Available Online

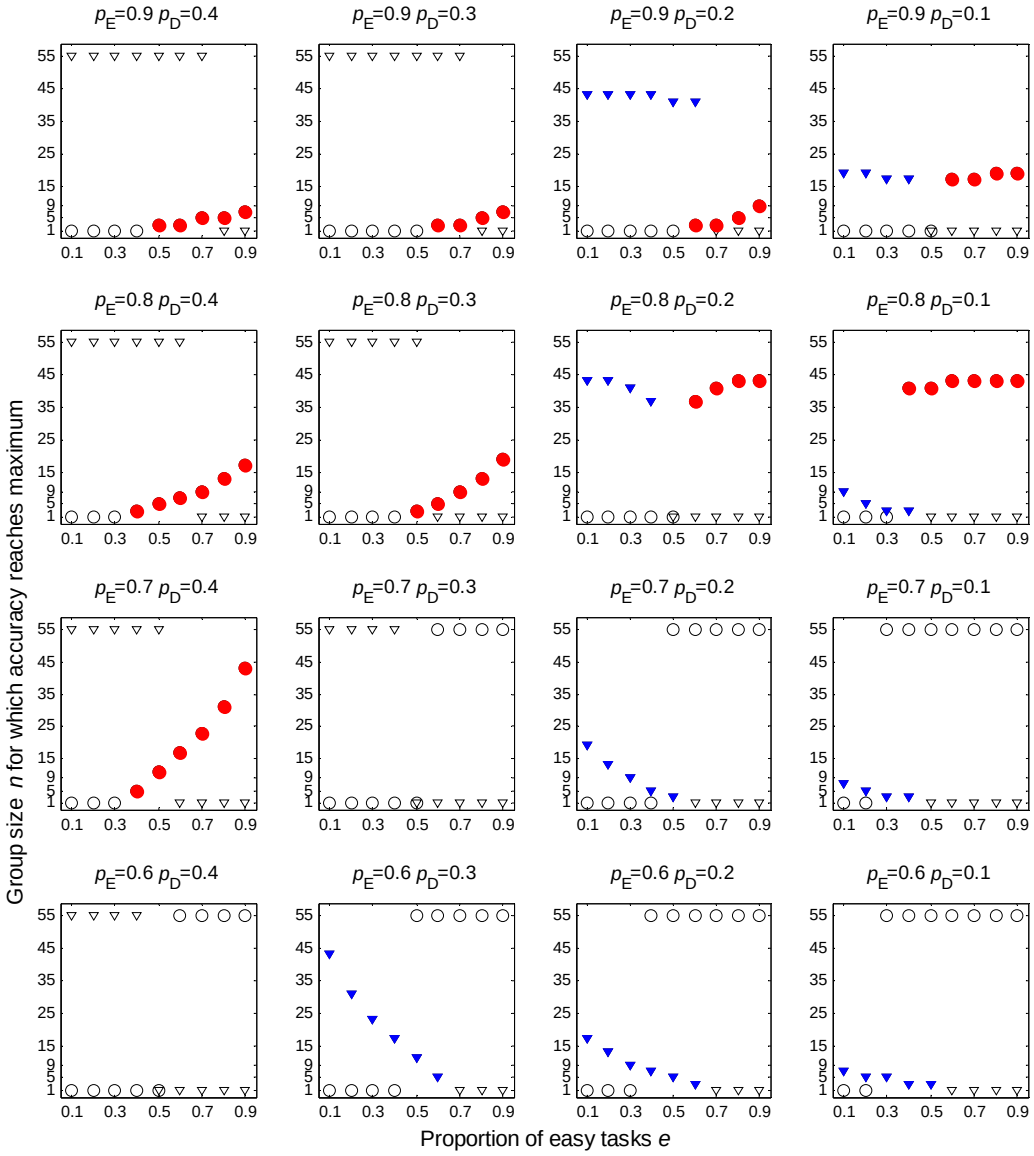

Figure S1. Group sizes for which group accuracy reaches maximum (circles) and minimum (triangles), for different combinations of task difficulties and proportions of easy tasks. Circles (triangles) in each panel show group size n at which group achieves maximum (minimum) accuracy, for sizes $1 \le n \le 55$. Full red circles (full blue triangles) denote cases where group reaches maximum (minimum) accuracy at group sizes that are larger than 1 but smaller than 55. Panels show results for different combinations of easy ($0.6 \le \overline{p}_E \le 0.9$) and difficult ($0.1 \le \overline{p}_D \le 0.4$) tasks. Each panel shows results for different proportions of easy tasks (x-axes, $0.1 \le e \le 0.9$). Panels above the diagonal represent friendly task environments, those in the diagonal neutral, and those below the diagonal unfriendly task environments.

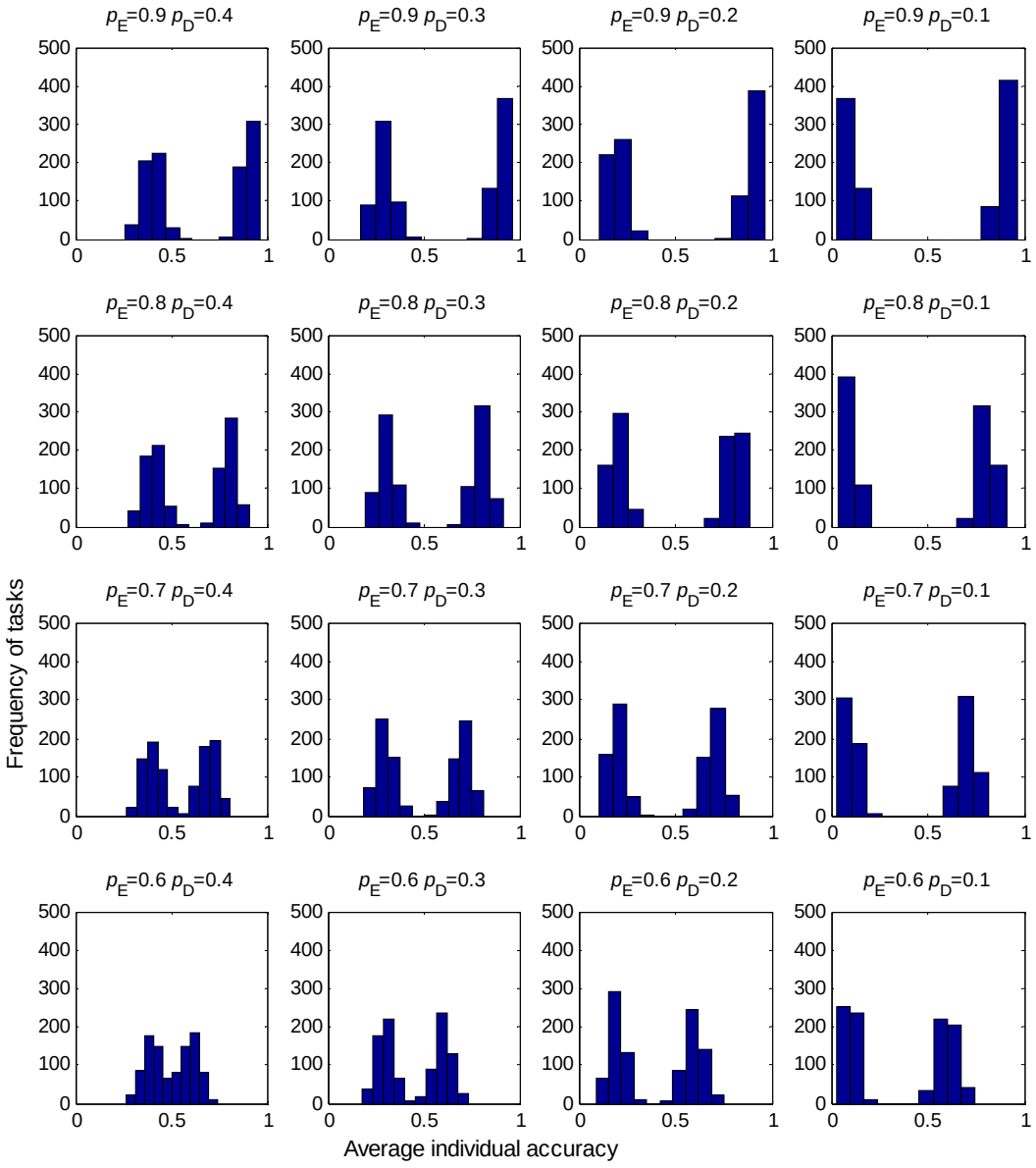

Figure S2A. Modeling more than two task difficulties. Easy tasks are drawn from distribution of task difficulties with mean $\overline{p}_{\mathrm{E}}$ and variance $\overline{p}_{\mathrm{E}}(1-\overline{p}_{\mathrm{E}})/(k+1)$. Difficult tasks are drawn from a beta distribution with mean $\overline{p}_{\mathrm{D}}$ and variance $\overline{p}_{\mathrm{D}}(1-\overline{p}_{\mathrm{D}})/(k+1)$. Here, variance is assumed to be small (k=100). There are total of 100 tasks. Each panel shows the resulting distribution of task difficulties.

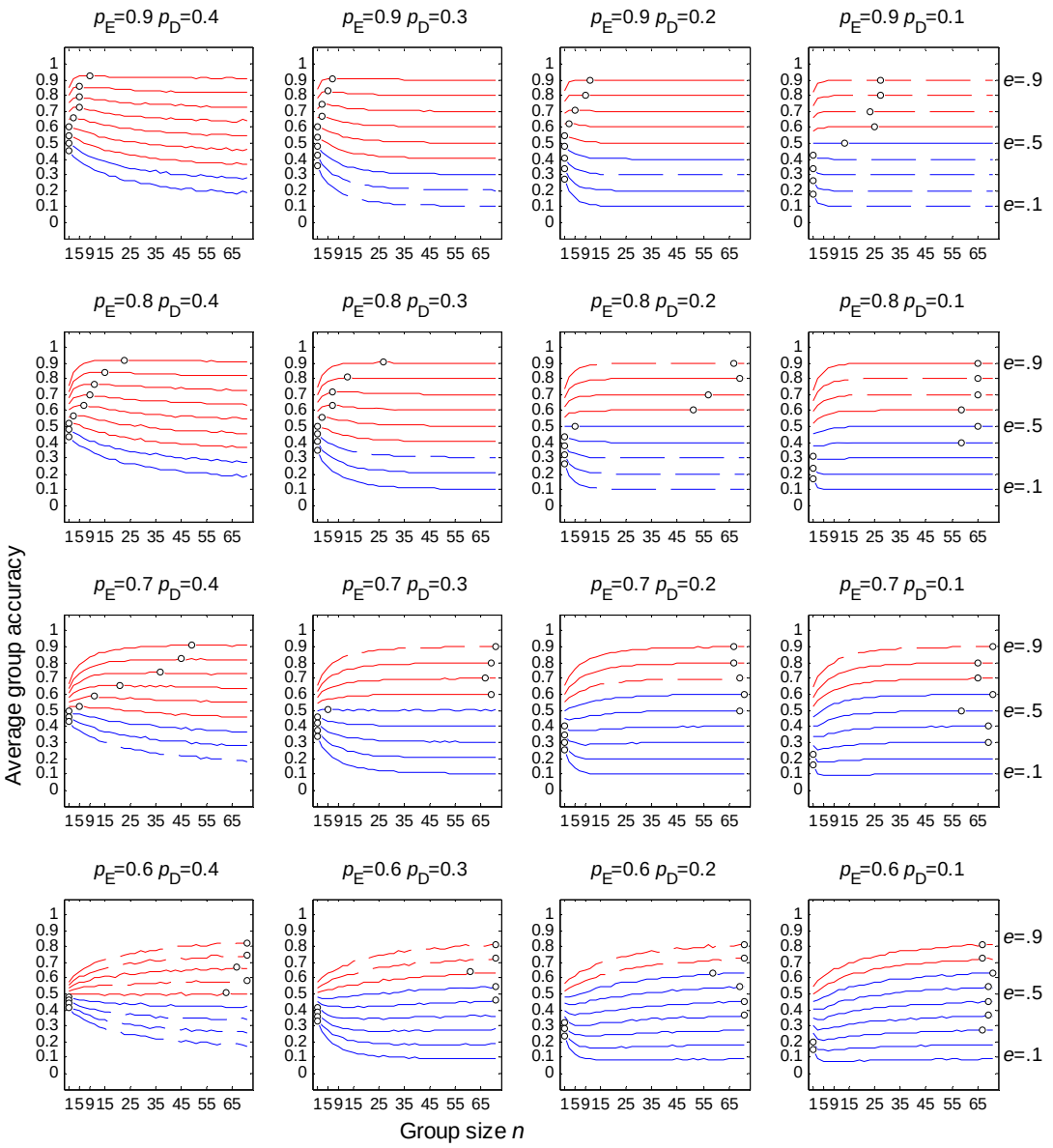

Figure S2B. Average group accuracy on tasks with difficulties displayed in the equivalent panels of Figure S2A. Variance is assumed to be small (k=100). Each panel shows changes in average group accuracy $\overline{P}$ as a function of group size *n*, different combinations of easy ($0.6 \leq \overline{p}_{\mathrm{E}} \leq 0.9$) and difficult ($0.1 \leq \overline{p}_{\mathrm{D}} \leq 0.4$) tasks, and different proportions of easy tasks ($0.1 \leq e \leq 0.9$). Red lines represent cases in which average individual accuracy across tasks $\overline{P}_{n=1} > 0.5$, blue lines are for $\overline{P}_{n=1} < 0.5$, and black lines for $\overline{P}_{n=1} = 0.5$, where $\overline{P}_{n=1}=e\overline{p}_{\mathrm{E}} + (1-e)\overline{p}_{\mathrm{D}}$. Circles show maximum value of $\overline{P}$ for each case. Dashed lines denote cases where $\overline{P}$ changes monotonically with *n* until it reaches *e*, while solid lines denote cases where $\overline{P}$ changes nonmonotonically, that is, reaches an upward or a downward peak at moderate group size *n* before reaching *e*. 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 (see main text for details).

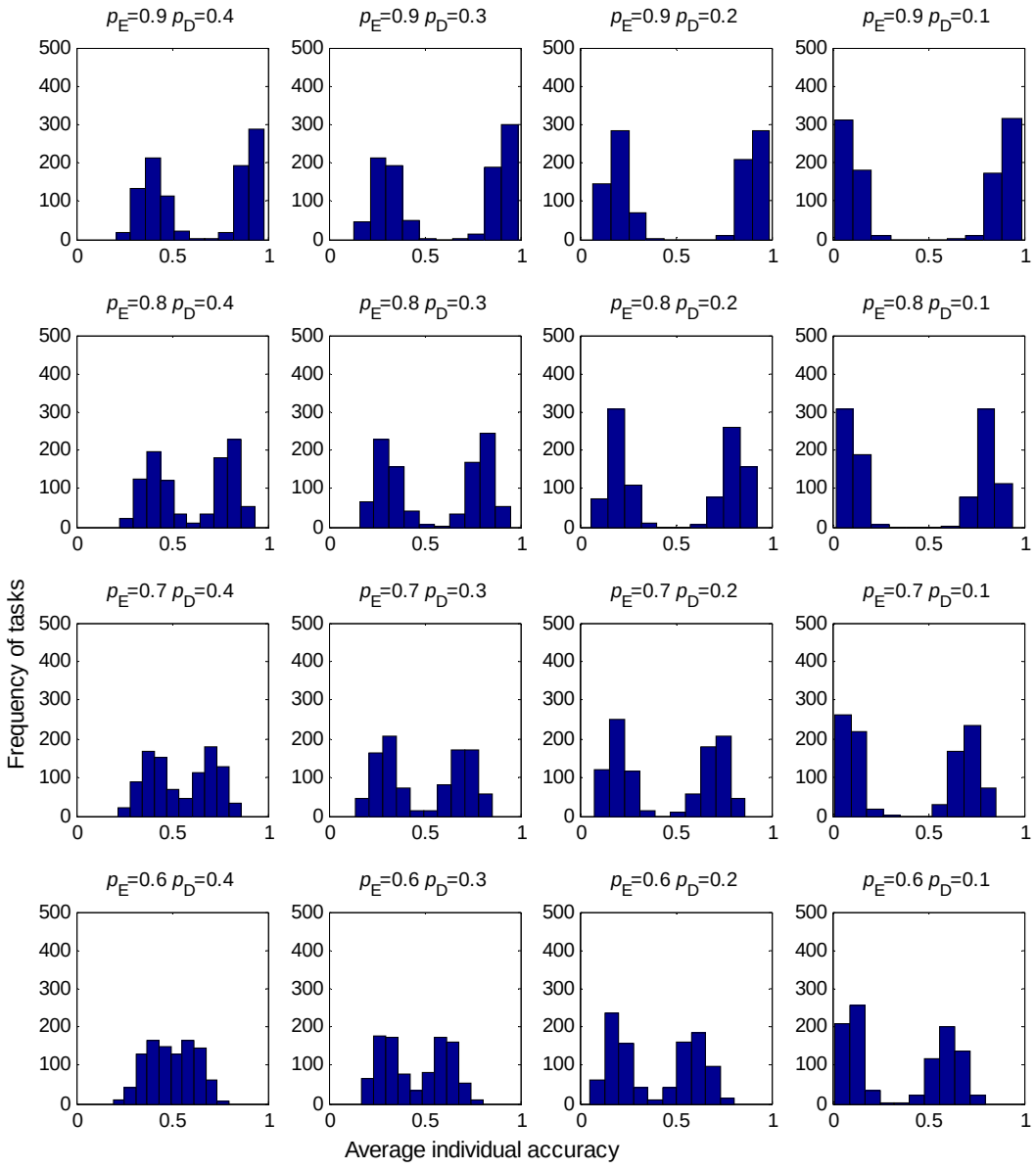

Figure S3A. Like Figure S2A, but variance of task difficulties is assumed to be moderate (k=50).

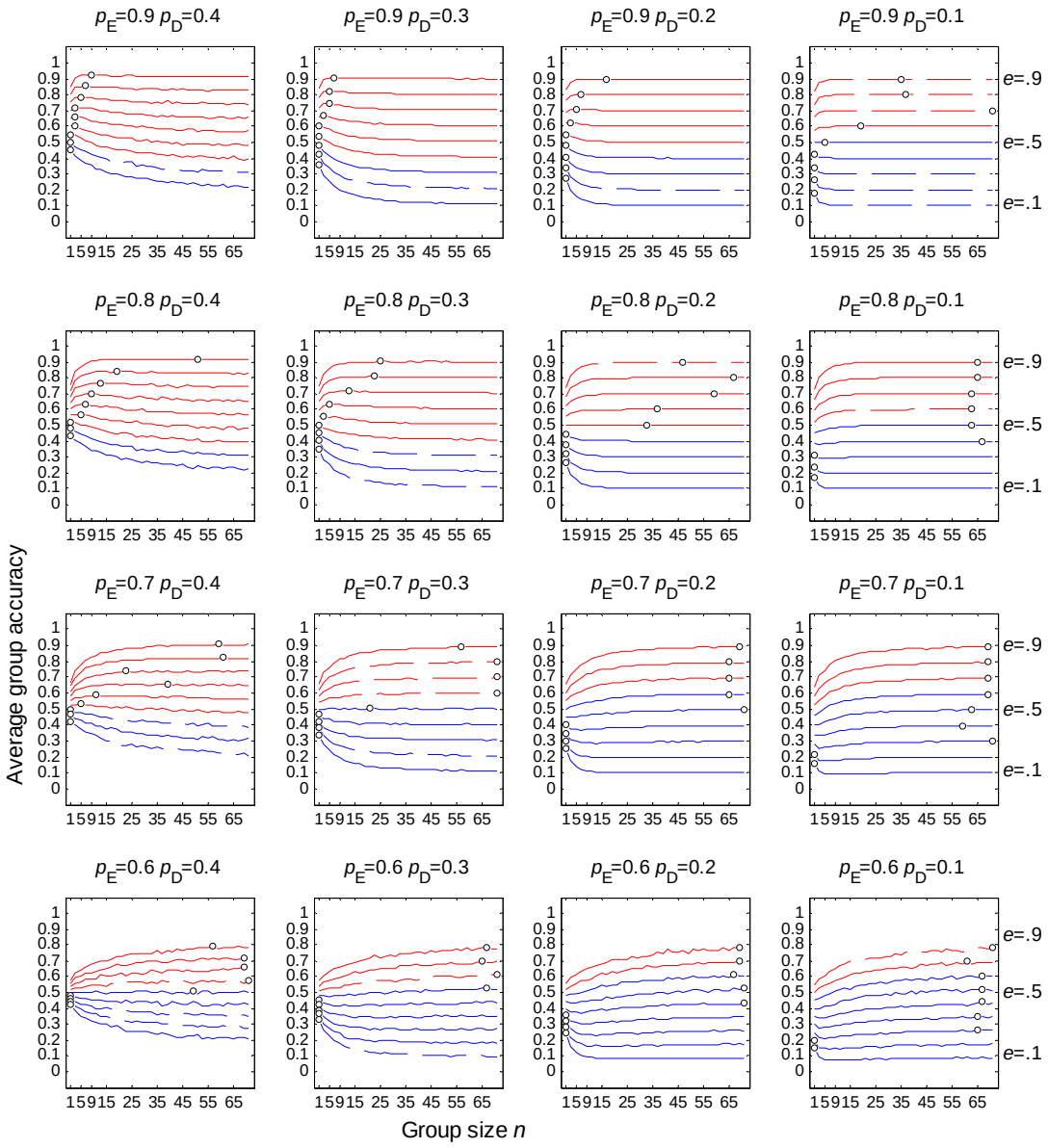

Figure S3B. Like Figure S2B, but variance of task difficulties is assumed to be moderate (k=50).

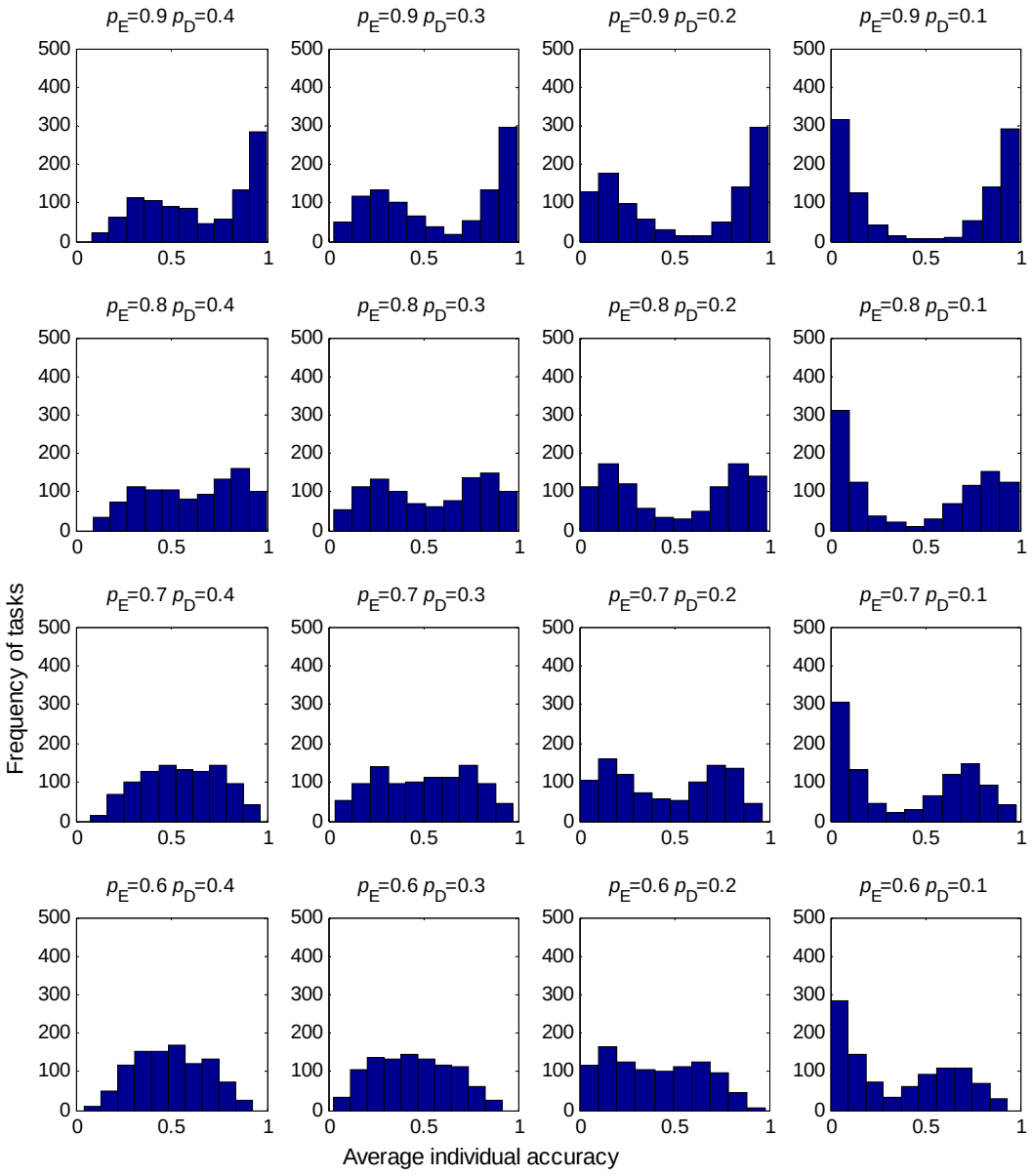

Figure S4A. Like Figure S2A, but variance of task difficulties is assumed to be large (k=10).

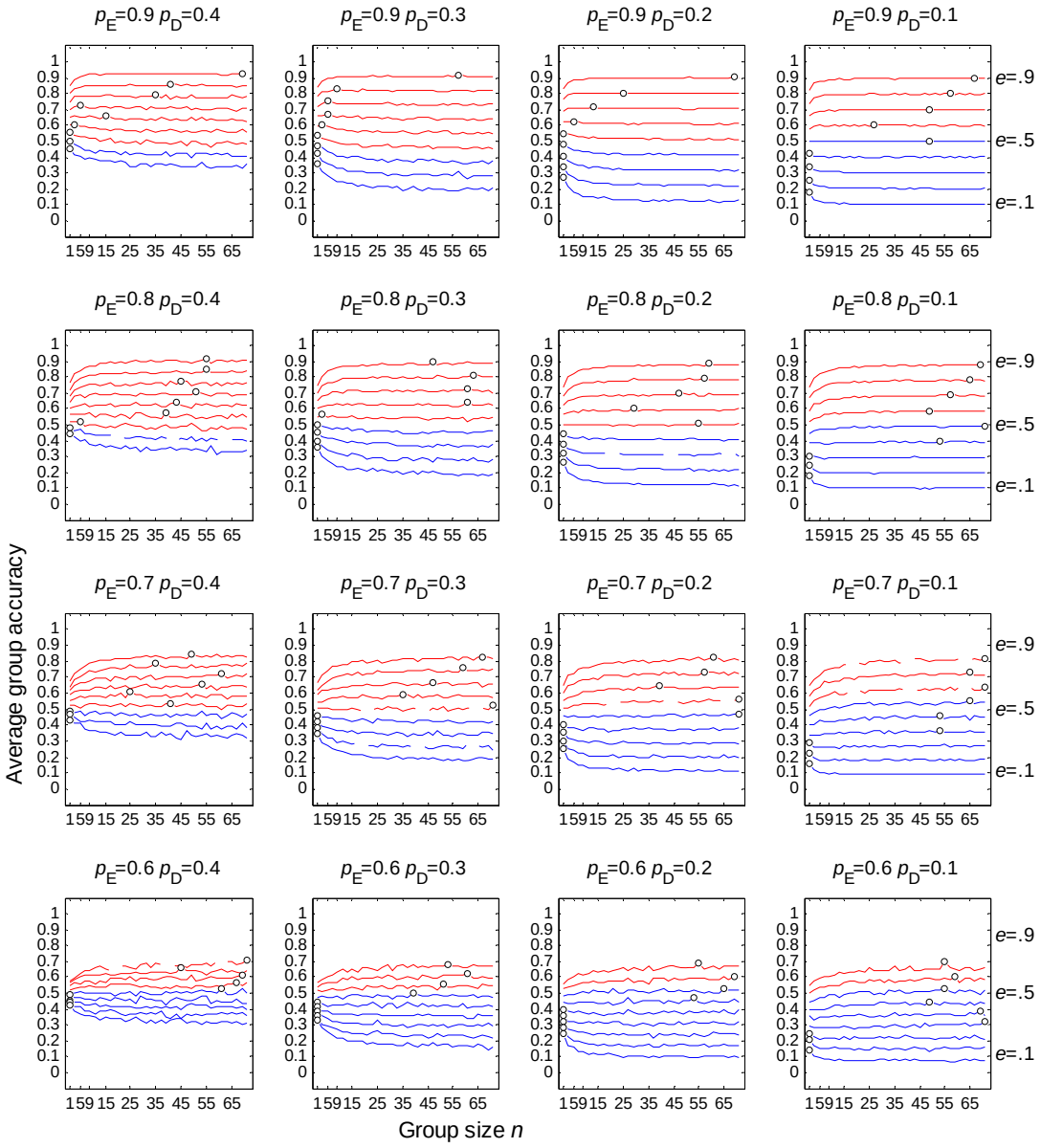

Figure S4B. Like Figure S2B, but variance of task difficulties is assumed to be large (k=10).

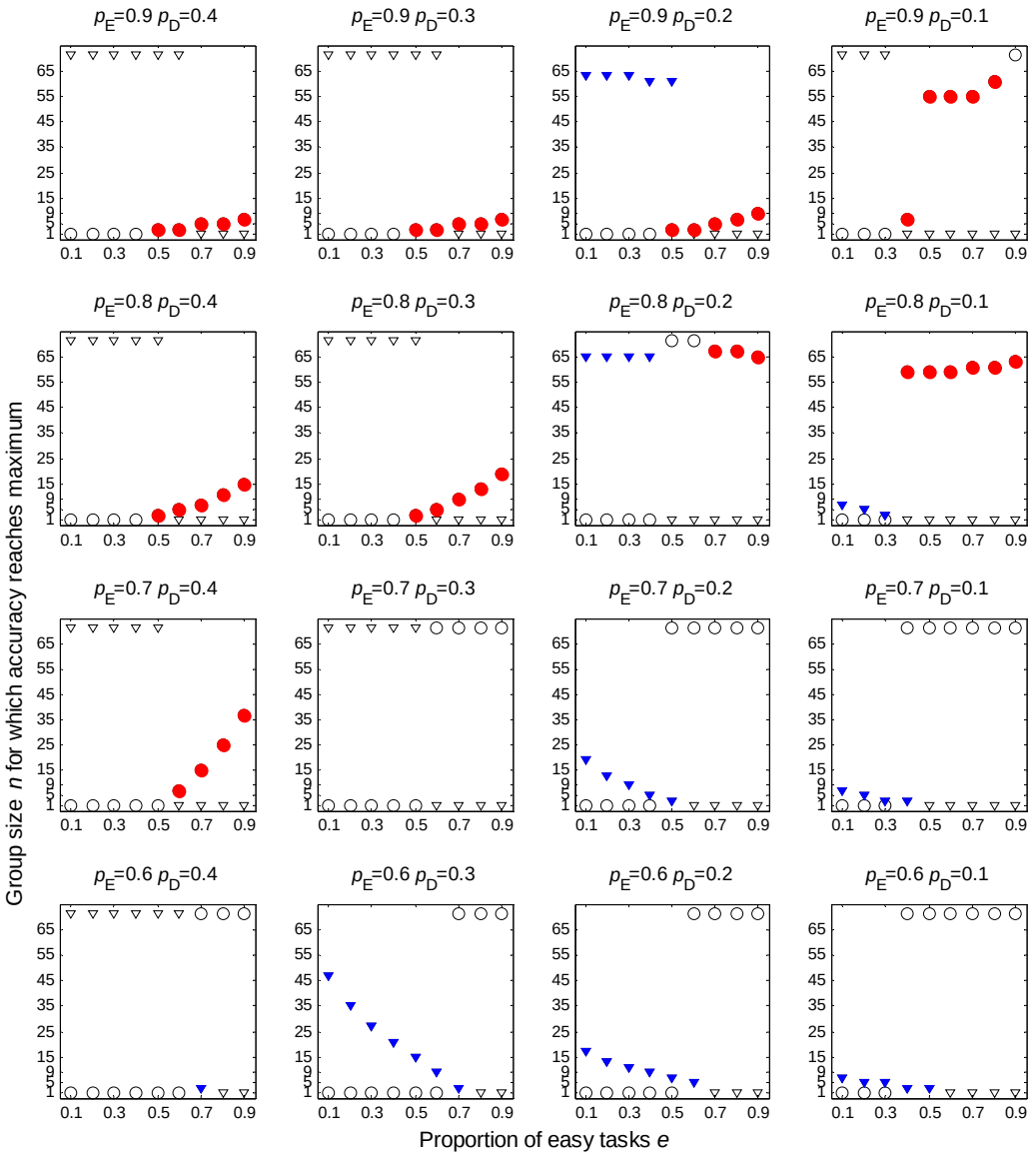

Figure S5. Results averaged across situations with different levels of correlations of individual judgments, for $0 \le r \le 1$: Group sizes for which group accuracy reaches maximum (circles) and minimum (triangles), for different combinations of task difficulties and proportions of easy tasks. Meanings of symbols are as in Figure S1. See Figure 3 for more detailed results.

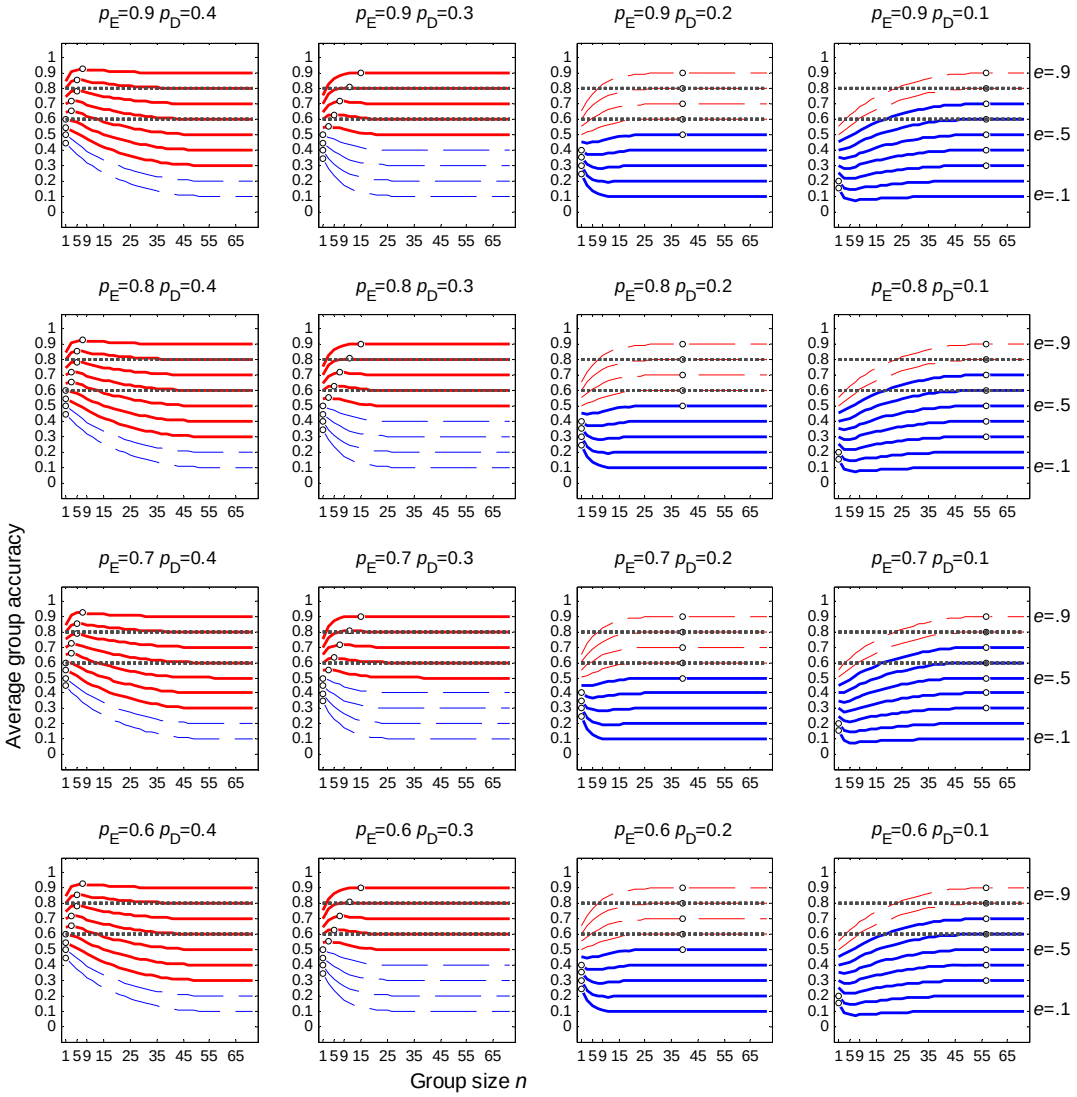

Figure S6. Sampling from finite populations without replacement accentuates previous results. Similar to Figure 2 in the main text, but group members are selected from a finite population of only N=71 members, without replacement. Instead of binomial, hypergeometric distribution is used to calculate accuracies of differently sized groups.

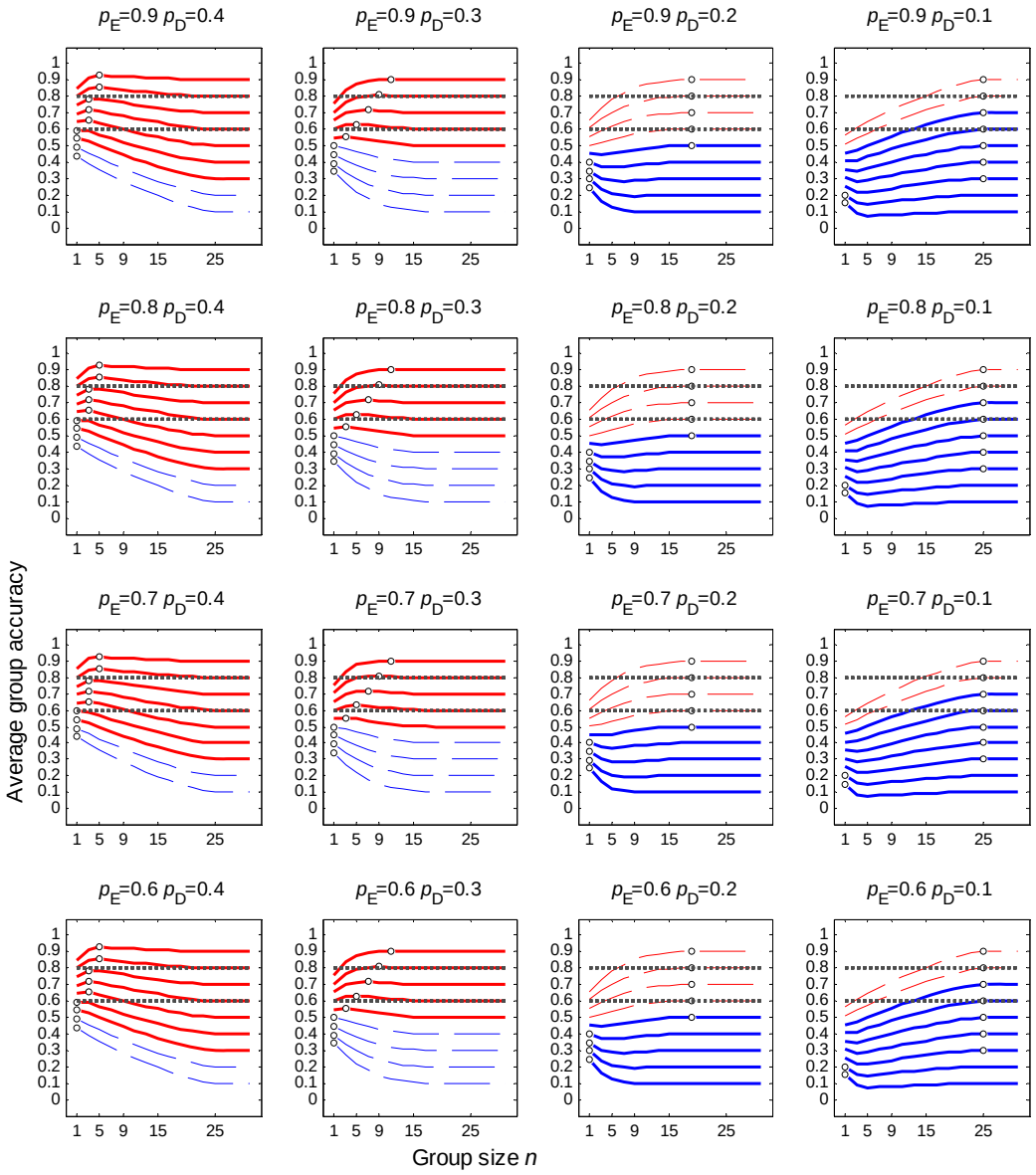

Figure S7. Sampling from finite populations without replacement accentuates previous results. Similar to Figure 2 in the main text, but group members are selected from a finite population of only N=31 members, without replacement. Instead of binomial, hypergeometric distribution is used to calculate accuracies of differently sized groups.