Crowd Wisdom Relies on Agents' Ability in Small Groups with a Voting Aggregation Rule

Marc Keuschnigg, Christian Ganser
Department of Sociology, Ludwig-Maximilians-University Munich, Germany
keuschnigg@lmu.de; christian.ganser@lmu.de

In the last decade interest in the "wisdom of crowds" effect has gained momentum in both organizational research and corporate practice. Crowd wisdom relies on the aggregation of independent judgments. The accuracy of a group's aggregate prediction rises with the number, ability, and diversity of its members. We investigate these variables' relative importance for collective prediction using agent-based simulation. We replicate the "diversity trumps ability" proposition for large groups, showing that samples of heterogeneous agents outperform same-sized homogeneous teams of high ability. In groups smaller than about 16 members, however, the effects of group composition depend on the social decision function employed: Diversity is key only in continuous estimation tasks (averaging) and much less important in discrete choice tasks (voting), in which agents' individual abilities remain crucial. Thus, strategies to improve collective decision-making must adapt to the predictive situation at hand.

Key words: averaging; combining judgments; diversity; social choice; voting

1. Introduction

The wisdom of crowds effect has drawn the attention of early scholars (e.g., Condorcet 1785, Galton 1907) and contemporary researchers alike (e.g., Hong and Page 2004, Lorenz et al. 2011) and, in the last decade, has grown highly popular in mainstream literature (Page 2007, Surowiecki 2004). The idea is simple: When predicting an unknown outcome (e.g., an object's weight or the future value of a financial product) the central tendency of a set of independent estimates represents the truth more closely than the typical individual estimation. Firms have thus been increasingly employing mechanisms of aggregating multiple opinions, especially when navigating markets which are difficult to predict (e.g., Bonabeau 2009, Soukhoroukova et al. 2012). We believe that combining independent judgments offers a valuable management tool to improve collective decision-making in both firms and public administration.

Our analysis identifies determinants of collective accuracy conditional on the aggregation mechanism employed. Using agent-based simulation we test the conditions under which judges' ability or diversity contribute more to the precision of aggregate prediction. Three key results emerge: (1)

Averaging is more effective than voting. (2) Adding diversity nearly always helps averaging more than adding ability. (3) In the case of plurality vote, however, ability remains crucial if groups are small. Depending on parameter values as well as on the error measure ability dominates collective accuracy in groups with up to 14–20 members.

In nominal groups the occurrence of crowd wisdom is a mathematical fact: With increasing size a collection of autonomous judges free of social interaction apart from some aggregation rule will almost always be more accurate than the expected value of a random draw from individual opinions (Hogarth 1978).[1] Variance of independent judgments is the key mediator for collective accuracy: With increasing diversity individual judgments are more likely to "bracket the truth" (Larrick and Soll 2006), permitting aggregation to cancel out contradictory biases. Closely related pieces of information instead narrow the variation of individual signals and undermine this effect (Lorenz et al. 2011).

Both mechanisms, the law of large numbers and the cancelation of contradicting errors, imply that when approximating a true value, a large and heterogeneous group outperforms samples of homogeneous experts—even when the crowd consists of error-prone individuals (Grofman et al. 1983, Hong and Page 2004). Gains from adding a less accurate but different judgment will be larger, as heterogeneity increases the effective sample size. Research on the wisdom of crowds thus capitalizes on the functional value of predictive diversity.[2]

We challenge this "diversity trumps ability" proposition in that it relies fundamentally on the particular aggregation algorithm employed to produce a group solution. While the proposition clearly holds for averaging across cardinal estimates (mean over N "evaluations" on a metric scale), its application to discrete choice situations (mode of N "votes" on a nominal scale) has not been studied systematically. Prior research has neglected how specific social decision functions moderate the accuracy effects of ability and diversity, and so the proposition's scope conditions are not yet fully understood. This blind spot is particularly surprising given group voting's ubiquity in both public administration and the private sector. Hence, our study also relates to the social choice literature (Balinski and Laraki 2010, Condorcet 1785), examining the information aggregation capacity of different social decision functions. Applications of both averaging and voting occur in a wide variety of contexts and, as will become clear, each social decision rule represents a distinct type of predictive task.

[1] The specific loss function (i.e., the operationalization of individual and collective error) limits generality: If all judges share the same bias (i.e., all over- or underestimating the criterion) and one uses mean absolute error the group average will equal the error of the average member. If the loss function is squared error, however, the group average will be more accurate than the average member even under identical individual bias. If at least one individual deviates from the crowd and falls on the other side of the truth averaging outperforms the average member for all convex loss functions (Larrick and Soll 2006).

[2] The mechanics of the wisdom of crowds effect under averaging are well-established at least since Hogarth (1978). Hogarth's formal analysis demonstrated that simply adding expertise to a predictive team will not optimize predictive accuracy unless the mean intercorrelation of individual judgments is reduced.

2. Simulation Approach

Brunswik's (1952) lens model, initially used to study clinical judgment (Hammond 1955), has been in more general use since the 1960s to describe human judgment in statistical terms (see Castellan 1973, Karelaia and Hogarth 2008). The model describes individual judgment as a linear combination of multiple cues providing probabilistic information on some criterion's value (see Figure 1). Although the lens model is unlikely to provide full understanding of human judgment, it permits evaluation of aggregation rules (see Hastie and Kameda 2005) and identification of corresponding determinants of collective accuracy.

Brunswik's framework allows taking into account a wide range of parameter values including the overall task environment, individual characteristics, and group features. The simulation consists of three components: a task environment, a pool of autonomous agents, and an aggregation algorithm.

1. The environment provides a criterion Q and a set of probabilistically related cues C_k ($k = 1, 2, ..., K$). The criterion reflects, for example, the future price of a stock, the market potential of an innovation, or the productivity of a job applicant. Cues might include a firm's profits, results from a market survey, or a resumé. The task entails individual selection of the superior alternative.

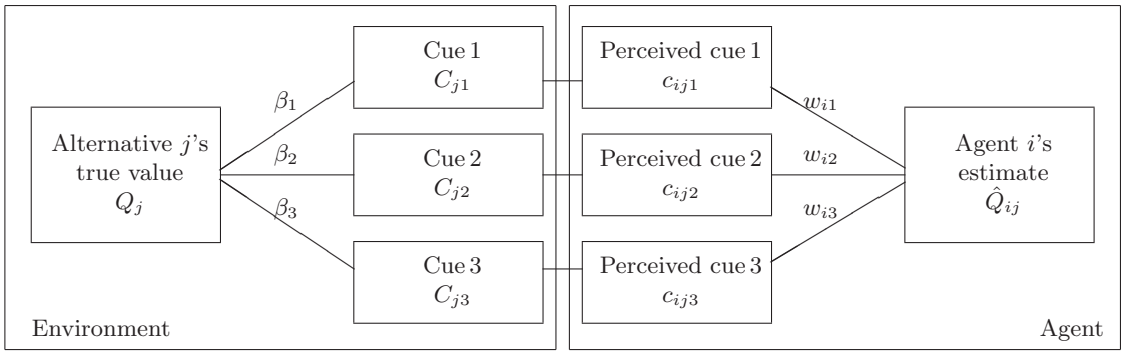
2. Agent $i = 1, 2, ..., N$ perceives cues and combines them into a subjective judgment $\hat{Q}_i$ providing an individual estimate of Q. We model agents as judges each combining a pair of perspectives and heuristics (Hong and Page 2004, 2012): Perpectives vary over agents as they differ in how precisely they perceive cue information. Heuristics describe individual weighting strategies for combining perceived cues. Both dimensions of predictive diversity limit positive correlation of individual judgments. Agents belong to a common type if both their perspectives and their heuristics are identical.

3. We create nominal groups of different size by randomly drawing individual judgments from the agent pool. We then apply two distinct social decision functions, averaging and plurality vote, in aggregating members' judgments into a single group solution. This allows us to study the effects of both task characteristics and group composition on collective accuracy, depending on the aggregation rule applied.

Our set-up closely follows that of Hastie and Kameda (2005). Suppose there are J stocks, innovations, or candidates to evaluate. Each alternative $j = 1, 2, ..., J$ holds a criterion value Q_j distributed normally across alternatives ($Q_j \sim \mathcal{N}(\mu, \sigma)$). Q_j is unobservable to agents but they receive noisy information from cues C_k which correlate with the criterion. In accordance with Hastie and Kameda (2005) we set $J = 10$ and $K = 3$.

$$Q_j = \beta_1 C_{j1} + \beta_2 C_{j2} + \beta_3 C_{j3} + u_j \tag{1}$$

Figure 1 **Lens model.**



Equation 1 captures ecological predictability, i.e. the degree to which one can predict the criterion based on cues. In our standard set-up, cues—each weighted by a factor β_k—correlate linearly with the criterion. The optimal function's model determination (R^2) provides a measure of task difficulty. Cue validity, for example, might be low in the case of financial investment but higher for a job interview. Consequently, buying a profitable stock should be more difficult than hiring a promising employee.

Besides task difficulty, both individual perception and weighting of cues determine an agent's predictive accuracy. Agents perceive cues with inter-individual variation. To capture differences in perspectives, we add an agent-specific perception error e_{ijk} to each cue value C_{jk} ($c_{ijk} = C_{jk} + e_{ijk}$). In our baseline world we randomly assign e_{ijk} from a normal distribution to each combination of agent, alternative, and cue ($N \times 10 \times 3$). We further define each agent's ability in perspectives as equaling the root mean squared perception error: $\text{RMSE}_{P_i} = \sqrt{\frac{1}{30}\sum_{j=1}^{10}\sum_{k=1}^{3} e_{ijk}^2}$. We chose the frequently used RMSE as a measure of differences between true and perceived cue values due to its useful property of penalizing strong deviations.

Agents combine perceived cues c_{ijk} in ways approximating the accurate weights β_k in the environmentally optimal function (equation 1). We assume agents to be knowledgeable about the linear relation in the environment. As agents differ in heuristics, however, they use various judgment policies to aggregate information. We randomly assign each agent some degree of individual misweighting, adding heuristic error v_{ik} to optimal weights β_k ($w_{ik} = \beta_k + v_{ik}$). Again, in the baseline world we draw v_{ik} from a normal distribution and randomly assign errors to each combination of agent and cue ($N \times 3$).[3] For each agent ability in heuristics equals RMSE of individual weights:

[3] If agents assign 0 weight to one or more cues they ignore the respective cue(s). Hence, our set-up considers the special case that judges use different cues.

$\text{RMSE}_{H_i} = \sqrt{\frac{1}{3}\sum_{k=1}^{3} v_{ik}^2}$. Note that $\text{cov}(e_{ijk}; v_{ik}) = 0$. Each agent i then estimates the criterion value Q_j of each alternative j:

$$\hat{Q}_{ij} = w_{i1}c_{ij1} + w_{i2}c_{ij2} + w_{i3}c_{ij3} \tag{2}$$

We randomly sample agents in groups g of different sizes $N = 4, 6, 8, ..., 16, 18, 20$. We specifically chose this range to compare aggregation rules in smaller groups. The range covers group sizes frequently observed in decision-making committees in both firms (e.g., management boards, work teams) and public administrations (e.g. cabinets, councils). The range includes group sizes 5 and 12 as studied by Hastie and Kameda (2005), and our ceiling of $N = 20$ corresponds to Hong's and Page's (2004). We increase group size in steps of two to economize on computing time.

3. Aggregation

The set-up permits direct comparison of aggregation rules. Following Hastie and Kameda (2005) we start with an estimation task as the fundamental judgment: Each agent provides a cardinal estimate of the criterion Q_j for each alternative j. We then translate the estimates into either a group mean (averaging) or a group vote for alternatives (plurality vote).

Averaging and voting relate to different types of predictive tasks. Accordingly, we conceptually distinguish between estimation and choice tasks. Estimation refers to predictions of a specific value from a metric scale. In an organizational perspective this could be a new product's expected sales volume, the likelihood of a partner's bankruptcy, or the expected price of raw material in one year's time. A choice task describes a situation of voting for a discrete alternative (e.g., category A, B, C, or D). Examples are the selection of a marketable technology from a set of innovations, the choice of trading partners, or identification of substitute inputs.

3.1. Averaging

Averaging considers group members' cardinal "evaluations" for each alternative and combines them into a single score. To process a group solution, each member i feeds individual estimates $\hat{Q}_{ij}$ for all alternatives j into aggregation. For each alternative j we then average the $\hat{Q}_{ij}$ over N group members ($\frac{1}{N}\sum_{i=1}^{N} \hat{Q}_{ij} = S_j$). The group solution is the alternative with the highest mean ($\max_j(S_j)$). In our standard set-up we then employ a binary loss function: Group g finds a correct solution $Y_g = 1$ if $\max_j(S_j) = \max_j(Q_j)$; otherwise $Y_g = 0$.

Averaging over individual estimates for each alternative j effectively neutralizes misjudgments: In the least beneficial cases of $\hat{Q}_{ij} > Q_j$ or $\hat{Q}_{ij} < Q_j$ for all i (i.e., all group members over- or underestimate the criterion), the group's average prediction S_j equals the expected value of a random draw from all individual judgments. Without "bracketing the truth" (Larrick and Soll

2006), a group's average prediction is always as accurate as the typical individual estimation (see footnote 1). However, if the $\hat{Q}_{ij}$ bracket Q_j (i.e., individual estimates spread to both sides of the truth), the group's average prediction must be more accurate than a random draw from individual judgments as $S_j - Q_j$ approaches 0.

Further, one can interpret each group average S_j as an estimate of the criterion's expected value $E(Q_j)$. Increasing group size results in S_j more closely distributed around the population mean. Corresponding to the law of large numbers collective accuracy increases with the square root of group size. If individual predictive errors are unsystematic and have 0 mean, $\lim_{N\to\infty}(S_j - Q_j) = 0$ holds.

Ability. In groups of high-ability members, individual estimates $\hat{Q}_{ij}$ will closely approximate each alternative's criterion value Q_j because individual error is small both in perception and in weighting. In relatively small groups, with limited error cancelation, ability is thus a crucial determinant of collective accuracy. Expertise, however, typically decreases the range of individual judgments (e.g., Hong and Page 2004, 2012). If experts closely approximate the criterion but exhibit similar biases, collective prediction cannot be fully accurate.

Diversity. When group members exhibit similar patterns of misjudgment, this prevents contradictory individual predictive errors (i.e. some individuals over- and others underestimating the true value) counterbalancing and cancelling out one another. "Bracketing" crucially depends on heterogeneity of individual estimates. Since negatively correlated predictive error increases the variation of individual estimations, heterogeneity must raise collective accuracy. At this point, our simulation reproduces the "diversity prediction theorem" (Page 2007), according to which collective squared error $(S_j - Q_j)^2$ equals average individual squared error $(\hat{Q}_{ij} - Q_j)^2$ minus the variance of members' estimates $(\hat{Q}_{ij} - S_j)^2$. The wisdom of crowds effect thus depends on both small individual error and diverse judgments.

3.2. Voting

Plurality rule only considers nominal "votes" for each member's first choice.[4] Each agent feeds one binary vote V_i into aggregation, indicating the alternative she evaluates most highly (if $\hat{Q}_{ij} = \max_j(\hat{Q}_{ij})$, then $V_i = j$). The mode of individual votes ($\mathrm{mod}_j(V_i)$) represents the group solution.

[4] As there are >2 alternatives, only judges' first preferences are considered. This relates to the fundamental problem of social choice: In the multiple options scenario ($J > 2$) attempts to consider the full preference list of individual voters (by pairwise comparison of alternatives) can, under specific conditions (Black 1958), lead to intransitive cycles on the aggregate level yielding no legitimate winner (Condorcet's paradox). Famously stated in Arrow's (1951) impossibility theorem, there is no democratic decision rule safely securing a transitive order of aggregated preferences. Recent work by Balinski and Laraki (2010), however, shows that eliciting individual preferences by cardinal "evaluation" of each alternative (rather than by ordinal pairwise comparison) circumvents many classical problems of plurality vote. We implicitly follow this road by testing evaluation-based averaging against comparison-based plurality vote. Yet, as we shall see below, we avoid pairwise comparison and disregard collective preference orders beyond the fact that voting aggregates less individual information than averaging.

Group g finds a correct solution $Y_g = 1$ if $\text{mod}_j(V_i) = \max_j(Q_j)$; otherwise $Y_g = 0$. If two (or more) alternatives receive an equal number of votes the group does not find the true alternative unequivocally. In our baseline world we code tied votes as incorrect outcomes.

Condorcet's (1785) jury theorem, originally stated for binary decisions (A or B), provides a starting point in formulating accuracy effects of both ability and diversity under plurality rule. In his classical contribution to social choice, Condorcet observed that, so long as each member of an electorate chooses the right alternative with probability $p > .5$, the majority of votes is bound to be correct. Moreover, with increases in the number of votes the probability of a correct collective prediction rapidly approaches 1: With rising numbers of judgments single misjudgments favoring the wrong alternative become less influential in turning the balance from the correct solution (cf., Conradt and List 2009, Grofman et al. 1983).

List and Goodin (2001) generalize Condorcet's jury theorem to the multiple options scenario without relying on a sequence of pairwise comparisons of alternatives (e.g., Arrow 1951). Most importantly, List and Goodin show that, unlike in the original "A or B" problem, to yield a correct collective outcome, plurality vote permits group members' abilities to choose the right alternative to be well below $p = .5$.

Ability. Competence is crucial to a group's gravity towards the truth. Being able to find the correct alternative is of particular relevance because—unlike averaging—plurality vote does not allow for direct cancelation of random error on the grounds of "bracketing the truth." For illustration consider a task environment with $J = 3$ alternatives and criterion values $Q_1 = 10$, $Q_2 = 8$, and $Q_3 = 2$. A group yields an incorrect solution if too many members underestimate the best alternative $Q_1 = 10$ while, at the same time, overestimating the value of an inferior alternative. Such a pattern of individual error raises the probability of collective misjudgment, favoring a wrong option over the correct solution. However, to yield a correct group solution each voter needs only choose the true alternative with probability $p > 1/J$ ($> 1/3$ in our example). In other words, individual probabilities of voting for the correct alternative must merely exceed the likelihood of voting for any of the remaining $J - 1$ wrong alternatives. As tipping the balance away from the correct alternative is more likely in cases of only a few independent judgments, ability should be particularly important in small groups.

Diversity. Again, heterogeneous judgments limit the correlation of individual errors. Under plurality vote, however, error cancelation occurs only insofar as misinformed judgments are less likely to cluster on any particular wrong option. By averting common patterns of misjudgment, diversity mainly contributes to clarity of collective choice: Clarity refers to how strongly a group prefers a modal category over all alternatives. Particularly in small groups—and unlike averaging—plurality vote is liable to produce tied outcomes. Reconsider our exemplary task environment with $J = 3$

alternatives. If too many votes cluster upon the second-best alternative $Q_2 = 8$, a group's solution becomes less clear—and, eventually, incorrect. Distributing incorrect choices away from the second-best alternative (i.e., randomization of misinformed choices across all alternatives Q_1, Q_2, and Q_3) requires substantial group size. Hence, diversity should be crucial only in large groups. Altogether, we expect diversity effects to be less pronounced than for the continuous estimation scenario.

4. Parameter Settings

For each group size we sample 100 random groups (without replacement) from an initial pool of 4 000 (200×20) agents. After generating 100 groups of a given size N, we replace all agents and draw 100 groups of size $N + 2$. We keep the environment's parameters and individual errors constant across group sizes. We repeat the whole sampling process 5 000 times. Over the runs we vary task difficulty and, for generalization, individual errors of perception and weighting: In each repetition we set new random values to generate the criterion Q_j, three related cues C_j, individual perception error e_{ijk}, and individual weighting error v_{ik}. The procedure provides us with 4.5 million groups (500 000 groups per size N).

In our standard set-up, cues are normally distributed with expected value 0 and standard deviation 20 ($C_k \sim \mathcal{N}(0, 20)$). We calculated the criterion Q_j from $.5C_1 + .3C_2 + .2C_3 + u$. For our baseline world we thus chose a compensatory weight dispersion (cf., Karelaia and Hogarth 2008: 407) of $\beta_1 = .5, \beta_2 = .3, \beta_3 = .2$. We increase the standard deviation of random error $u \sim \mathcal{N}(0, \sigma)$ over repetitions, such that task difficulty increases as R^2 of the environmental function declines from nearly 1 to .53 (median predictability $R^2 = .80$). We reverse and z-standardize R^2 to get a measure of task difficulty (high values represent relatively hard tasks).

Although we use simulated agents to process group solutions, the experimental design should represent, as closely as possible, common natural environments of human decision-making. We thus set parameter ranges at plausible real world intervals (see Karelaia and Hogarth 2008 for reference). However, as in every simulation study, parameter values are ultimately arbitrary and one may not interpret them in absolute terms but only in relation to each other.

To calibrate our manipulation range of ability, we consider agents to have some idea about the range of possible criterion values. Technically this implies that individual errors of perception (e_{ijk}) and weighting (v_{ik}) have standard deviations not exceeding the standard deviation of C_k and the true values of β_k respectively. Hence, we calculate perceived cues from $C_{jk} + e_{ijk}$ with $e_{ijk} \sim \mathcal{N}(0, 15)$. Similarly, we chose standard deviations of weighting errors v_{ik} slightly lower than the values of true weights: Agents combine perceived cues using individual weights $w_{ik} = \beta_k + v_{ik}$, with v_{ik} normally distributed with $\mu = 0$ and $\sigma(v_{i1}) = .35, \sigma(v_{i2}) = .20, \sigma(v_{i3}) = .15$.

With this calibration of our simulation's behavioral core we closely reproduce findings on real human judgment: The mean correlation between the criterion Q_j and individual estimates $\hat{Q}_{ij}$ is .57. In a meta-analysis of 86 experimental studies in the lens model tradition, Karelaia and Hogarth (2008) report a mean achievement of .56 (while mean environmental predictability in these studies is .81). When considering 1 000 environments with a mid-level task difficulty (optimal function's $R^2 = .80$), on average 38% of our agents find the correct solution autonomously. If we further increased the ability range, the share of correct individuals would drop considerably. In this case voting—which primarily rests on individual accuracy—would be a very weak aggregation rule altogether.[5]

Within each group g we measure ability in two dimensions: (1) Ability in perspectives equals members' average perception error $\text{RMSE}_{P_g} = \frac{1}{N}\sum_{i=1}^{N}\text{RMSE}_{P_i}$. (2) Ability in heuristics is the members' average weighting error $\text{RMSE}_{H_g} = \frac{1}{N}\sum_{i=1}^{N}\text{RMSE}_{H_i}$. Since each agent combines ability in perception with ability in weighting, we jointly analyze predictive error by pooling both variables $(\text{RMSE}_{P_g} + \text{RMSE}_{H_g})$ into a single indicator of the group's g ability A_g.

Likewise, we collect two indicators of group diversity: (1) To measure diversity in perspectives we calculate pairwise correlations ρ_i^P of perception errors e_{ijk} over group g's N members. The mean intercorrelation of individual errors $\bar{\rho}_g^P = \sum_{i=1}^{N}\rho_i^P/N$ provides a group-level measure of perceptive diversity. (2) Similarly, we measure diversity in heuristics as the mean $\bar{\rho}_g^H = \sum_{i=1}^{N}\rho_i^H/N$ of all pairwise correlations of agents' weighting errors v_{ik} within each group. We allow both diversity measures to range from –1 to 1. We chose the correlation coefficient as dissimilarity measure because the diversity effect depends upon relative direction of errors. Our operationalization is thus consistent with the underlying theoretical concept. Again, we combine a group's diversities in perspectives and in heuristics $(\bar{\rho}_g^P + \bar{\rho}_g^H)$ into a single measure of diversity D_g. Both variables, ability and diversity, are reversed (i.e., high values represent high ability and high diversity respectively) and z-standardized (mean 0, standard deviation 1) for relative interpretation.[6]

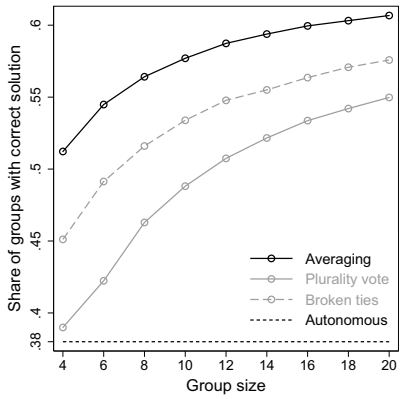
[5] In our sensitivity analysis (see Appendix A2, "accuracy of judges") we manipulate the range of ability to demonstrate how alternative parameter values affect our results. We relax the assumption of smart judges and make agents less informed about the range of possible criterion values. We do this by increasing the standard deviation of perception errors e_{ijk} to 20 and the standard deviation of weighting errors v_{ik} to .5, .3, and .2. Making judges less accurate strengthens the effect of ability on collective accuracy. If we further increased the range of ability, individual expertise would become an absolute necessity for collective accuracy under voting. Hence, our set-up evaluates the "diversity trumps ability" proposition under conservative testing conditions.

[6] Pooling does not affect our findings. Our main results appear similarly in a separated analysis for perspectives and heuristics (see Appendix A1). Unlike prior accounts (e.g., Hogarth 1978, Karelaia and Hogarth 2008) we refrain from simpler measures of ability and diversity such as the mean correlation of members' predictions with the criterion (for a group's ability) and the mean intercorrelation of individual predictions (for a group's diversity) because these measures are collinear: If members' predictions strongly associate with the criterion (i.e., members are highly able) mean pairwise correlation of predictions is inherently high (i.e., the group is less diverse). Surely, this would reflect reality as teams of experts are necessarily more homogeneous than groups of laymen (e.g., Hong and Page 2004, 2012). For our analytical purpose, however, we give preference to independent variation of both concepts.

5. Results

Figure 2 plots the share of correct group solutions ($Y_g = 1$) against group size for both averaging and plurality vote. Both group size and social decision function are crucial moderators of collective accuracy: First, we replicate the well-known result that size increases collective accuracy (cf., Grofman et al. 1983, Hogarth 1978). As compared to autonomous judgment, where 38% of agents find the true alternative, averaging (plurality vote) accounts for 58% (49%) of correct choices across all group sizes. Second, it shows that averaging clearly outperforms plurality vote (see solid lines in Figure 2). As theoretically expected, this superiority reflects the aggregation of more individual information under averaging. Only if we remedy a crucial inefficiency of plurality rule and break tied votes at random (see dashed line) we are able to approximate the "robust beauty of majority rules" (Hastie and Kameda 2005: 505).[7] Randomization, however, lacks legitimacy as a social decision rule and is hardly accepted in most modern-day administrations and corporate boards (e.g., Elster 1989). In our standard set-up we therefore code tied votes as incorrect choices (see our sensitivity analysis in Appendix A2 for a relaxation). Third, size is more important under plurality vote. We will see that voting requires more members to bring about a positive diversity effect on collective accuracy.

Figure 2 Group Size and Collective Accuracy.

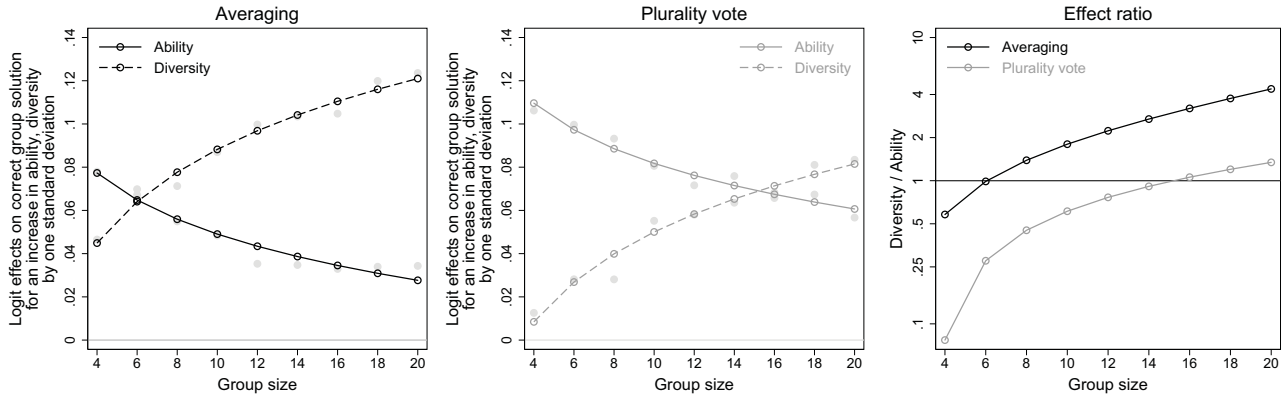


To assess the relevance of both ability A_g and diversity D_g for collective prediction, we estimate a binary logistic regression on the probability of a correct group solution for each group size N and aggregation mechanism (9×2 models):

$$\log\left(\frac{Pr(Y_g = 1)}{1 - Pr(Y_g = 1)}\right) = \hat{\beta}_0 + \hat{\beta}_A A_g + \hat{\beta}_D D_g \tag{3}$$

[7] Under plurality rule 15.3% of groups yield no legitimate winner due to tied votes. In tied groups the mean number of alternatives receiving equal votes is 2.4. Randomization endows 28.1% of these groups a correct solution.

Figure 3 Effects of Ability and Diversity on Collective Accuracy.



with $g = 1, 2, ..., 500\,000$ for each N. Figure 3, panel left and center, summarizes the results. The y-axis indicates estimated changes in the left hand side of equation 3 following an increase of ability or diversity by one standard deviation. Each vertical pair of gray dots represents these logit effects of ability and diversity for a given group size. In the left panel we apply our simulation approach to the well-documented task environment of continuous estimation (e.g., Galton 1907, Hogarth 1978). To smooth random estimation error we regress $\beta_A = \hat{\beta}_0 + \hat{\beta}_1 \log N + \hat{u}$ (see solid line) and $\beta_D = \hat{\beta}_0 + \hat{\beta}_1 \log N + \hat{u}$ (see dashed line). R^2 is .93 and .98 respectively. The middle panel shows a similar analysis for discrete choice.

The right panel in Figure 3 combines all information based on smoothed estimates. To visualize the relative importance of expertise vs. heterogeneity the y-axis displays the ratio of logit effects from diversity and ability ($\hat{\beta}_D / \hat{\beta}_A$) on group performance. We log the y-axis to make the relative comparison of diversity to ability symmetric. A ratio of 1 means equal benefits; values > 1 (< 1) indicate larger effects from diversity (ability). A ratio of 2 (.5), for example, means that diversity (ability) is twice as important as ability (diversity) for collective accuracy.

Under averaging, except for very small groups of up to 6 members, the effect ratio of diversity and ability is well above 1. Validating our testbed, we find clear support for the well-documented relevance of diversity under averaging (cf., Hogarth 1978, Hong and Page 2004, Larrick and Soll 2006). Under plurality vote, however, diverse groups outperform highly able groups only when group size is sufficiently large. In groups smaller than 16 members, ability is more important than diversity for collective accuracy.

Our main finding is clearly in line with theoretical considerations implying that individual competence is crucial for a small electorate's gravity towards the truth (Condorcet 1785, Grofman et al. 1983, List and Goodin 2001). The analogy of "bracketing the truth" in discrete choice, i.e. the spreading of misdirected votes away from a particular wrong alternative, relies on a larger number of individual votes. Diversity effects are thus substantial only in larger electorates.

Table 1 corroborates our results, reporting average marginal effects (AME) on the probability of a correct group solution for both averaging and plurality vote.[8] For each aggregation mechanism we estimated one binary logistic regression, taking into account ability A_g, diversity D_g, and task difficulty T as well as group size N_g as a moderator (all regressors are z-standardized):

$$\log\left(\frac{Pr(Y_g=1)}{1-Pr(Y_g=1)}\right) = \hat{\beta}_0 + \hat{\beta}_1 A_g + \hat{\beta}_2 D_g + \hat{\beta}_3 T + \hat{\beta}_4 \log N_g + \hat{\beta}_5(A_g \times \log N_g) + \hat{\beta}_6(D_g \times \log N_g) + \hat{\beta}_7(T \times \log N_g) \quad (4)$$

Table 1 Average Marginal Effects (AME) on Correct Group Solution.

	Averaging			Plurality vote		
Group size	4–8	10–14	16–20	4–8	10–14	16–20
Ability	1.163	.732	.496	1.673	1.344	1.126
Diversity	1.096	1.636	1.913	.454	.985	1.127
Task difficulty	–8.645	–9.997	–10.645	–7.095	–8.743	–9.493
log Group size	3.007	2.870	2.770	5.252	5.214	5.148

Logistic regressions on probability of correct group solution ($Y_g = 1$); z-standardized regressors; AME×100 reported.

To illustrate interaction with size, we display AME for three intervals of group size, 4–8, 10–14, and 16–20. Apparently, cue validity is an important predictor of group accuracy. Increasing task difficulty by one standard deviation lowers the probability of correct group solution across group sizes by 9.8% (averaging) and 8.4% (plurality vote) respectively. Task difficulty is particularly influential in large groups, as small groups are already more error-prone regardless of cue validity. Likewise, task difficulty plays a larger role under averaging, as voting leads to less accurate collective predictions overall. In line with Figure 2, increased group size fosters accuracy, particularly under plurality rule.

Compared to the grave consequences of both group size and task difficulty, absolute effects of ability and diversity remain rather small. This, however, reflects our study's design and does not constrain our findings' relevance. In terms of relative interpretation, diversity clearly outstrips ability in improving collective accuracy under averaging. In the case of plurality vote, however, ability remains crucial in smaller groups.

To identify the boundary conditions of our finding, we conducted a series of robustness analyses (see Appendix A2). Altogether, variations in parameters and the specific measurement of group performance do affect our estimates. Most pronounced, if task difficulty is high and judges' expertise is low, ability determines the performance of voting groups well beyond 20 individuals. Still, moderating effects are limited and do not impinge our main finding: In small groups accuracy of plurality vote relies on individual competence rather than predictive diversity.

[8] As our total sample size is artificial, we do not interpret inferential statistics.

6. Implications

Our generalizations are limited to nominal groups exhibiting autonomous decision-making and substitution of social interaction by an aggregation algorithm. Still, our study provides a necessary baseline in evaluating decision-making in real human teams, permitting study of fundamental group processes isolated from social confounders. The positive effect of group size on collective accuracy demonstrated a general advantage of nominal groups over individuals in making complex predictions. This result conforms to the law of large numbers and is scarcely new (Galton 1907, Grofman et al. 1983, Hogarth 1978). Similarly, we replicate the finding that averaging leads to more accurate group prediction than voting (Balinski and Laraki 2010). However, our study reveals several additional implications.

We varied both ability and diversity of groups by assigning individuals different cue-perception and cue-weighting schemes. Controlling for contextual factors such as group size and task difficulty we were able to show that the social decision function is an important moderator of accuracy effects due either to ability or to diversity.

Under averaging diversity leads to increased variation of individual judgments and thus enhances the cancelation of individual errors. At the same time individuals' ability is hardly a leverage to collective accuracy in continuous estimation tasks. Hence, when aggregating cardinal "evaluations" on a metric scale, adding diversity helps collective accuracy more than adding ability. In discrete choice tasks, however, diversity is less relevant, as voting limits error cancelation. By fostering modal clustering at the true category, ability has greater influence upon group accuracy in small groups. If tasks are difficult and judges' expertise is low, ability dominates the performance of electorates well beyond the size of 20.

The current analysis pits ability against diversity but measures each *post hoc*. Our findings will thus be useful in assembling real predictive teams only prospectively when ability and diversity are observable *ex ante*. Still, our findings are relevant for organizations seeking to implement applications of collective prediction: Combining our results, we argue that implementations of collective prediction require careful design. Distinctions both between predictions of metric values and discrete categories and between small and large groups proved highly relevant. Strategies resting primarily on maximizing diversity—as Hong's and Page's (2004, 2012) work suggests—are premature, as members' competence is important in the case of voting in small groups.

Consequently, one must calibrate group characteristics to the specific predictive problem at hand: First, whenever possible, group solutions should be sought under averaging rather than voting. Cardinal "evaluation" considers more individual information than ordinal "comparison," facilitates increased error cancelation, and circumvents equivocal outcomes. Second, large groups prove substantially more precise than small juries. Adding judges pays off, regardless of the type

of task or aggregation mechanism applied. Third, most electorates benefit from diverse rather than highly competent members. Predictive heterogeneity limits the correlation of individual errors and thus increases effective sample size. Most important for our study, however: in determining collective accuracy, diversity is crucial only in large groups and/or in cases of aggregation via averaging. Hence, if forced to plurality vote in a small group—which is often the case in decision-making committees in both firms and public administrations—the electorate must contain highly competent individuals.

Acknowledgments

We thank Norman Braun, Stefan Klößner, Volker Ludwig, Heiko Rauhut, Jan Schikora, three anonymous reviewers, and the editors for helpful comments. The article also greatly benefited from discussions at several workshops and conferences including the session 'Models of Innovation, Adaptation, and the Wisdom of Crowds' at the 2013 Congress of the Swiss Sociological Association in Bern and the 'Forschungskolloquium Empirie' at the Institute of Sociology in Bern.

Appendix

A1 Perspectives and Heuristics

Figures 4 and 5 summarize our findings from a separated analysis of perspectives and heuristics. Our key result shows similarly to our joint analysis of perception and weighting errors: Diversity matters more under averaging; under plurality vote, however, ability remains crucial for groups smaller than 14–18 members.

The separated analysis also illustrates differences in relative strengths of accuracy effects from ability and diversity regarding perception and weighting. These differences, however, are due to our design: Ability appears to be of higher relevance for heuristics than for perspectives, as low weighting ability can reverse the sign of weights. This is not so in the case of ability in perception. Consequently, for heuristics we see a slower increase in the effect ratio of diversity and ability against group size.

Figure 4 Ability and Diversity in Perspectives.

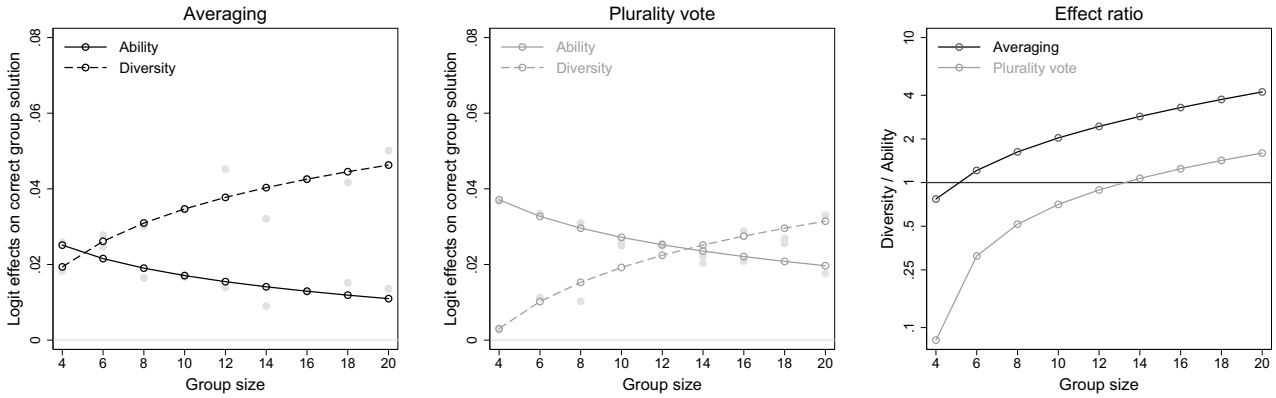
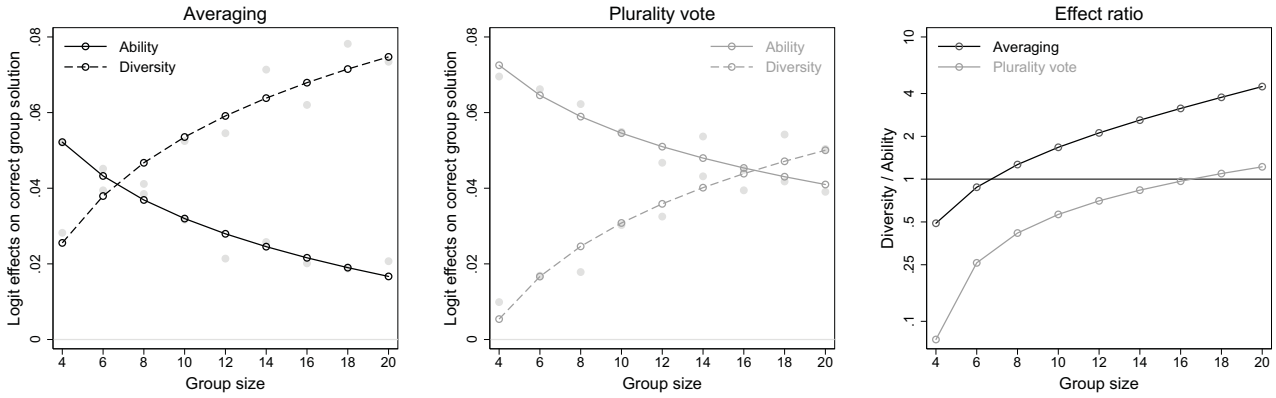


Figure 5 Ability and Diversity in Heuristics.



A2 Sensitivity Analysis

To generalize our key result, we conducted the following robustness analyses (see Figure 6; cf., Karelaia and Hogarth 2008 for a discussion of potential moderators of collective judgment):

Task difficulty. We re-ran our simulation with median environmental predictability $R^2 = .90$ (simple task) and $R^2 = .70$ (difficult task). Unlike averaging, plurality vote is sensitive to variations in task difficulty such that ability becomes more influential with decreasing cue validity.

Number of alternatives. Similarly, the number of alternatives can be interpreted as a measure of task difficulty. We tested $J = 6$ (simple task) and $J = 20$ (difficult task). Again, substantial differences occur only for voting where ability's relevance increases with the number of alternatives.

Accuracy of judges. Increasing the manipulation range of individual errors of perception ($\sigma(e_{ijk}) = 20$) and weighting ($\sigma(v_{i1}) = .50, \sigma(v_{i2}) = .30, \sigma(v_{i3}) = .20$) sharply defers the effect ratio's crossing of 1 for both aggregation mechanisms (see also footnote 5). Under voting, which relies more strongly on judges' accuracy, diversity effects dominate ability effects only in groups of 20 and more members. Higher mean accuracy ($\sigma(e_{ijk}) = 10; \sigma(v_{i1}) = .25, \sigma(v_{i2}) = .15, \sigma(v_{i3}) = .10$) instead limits the relative importance of ability.

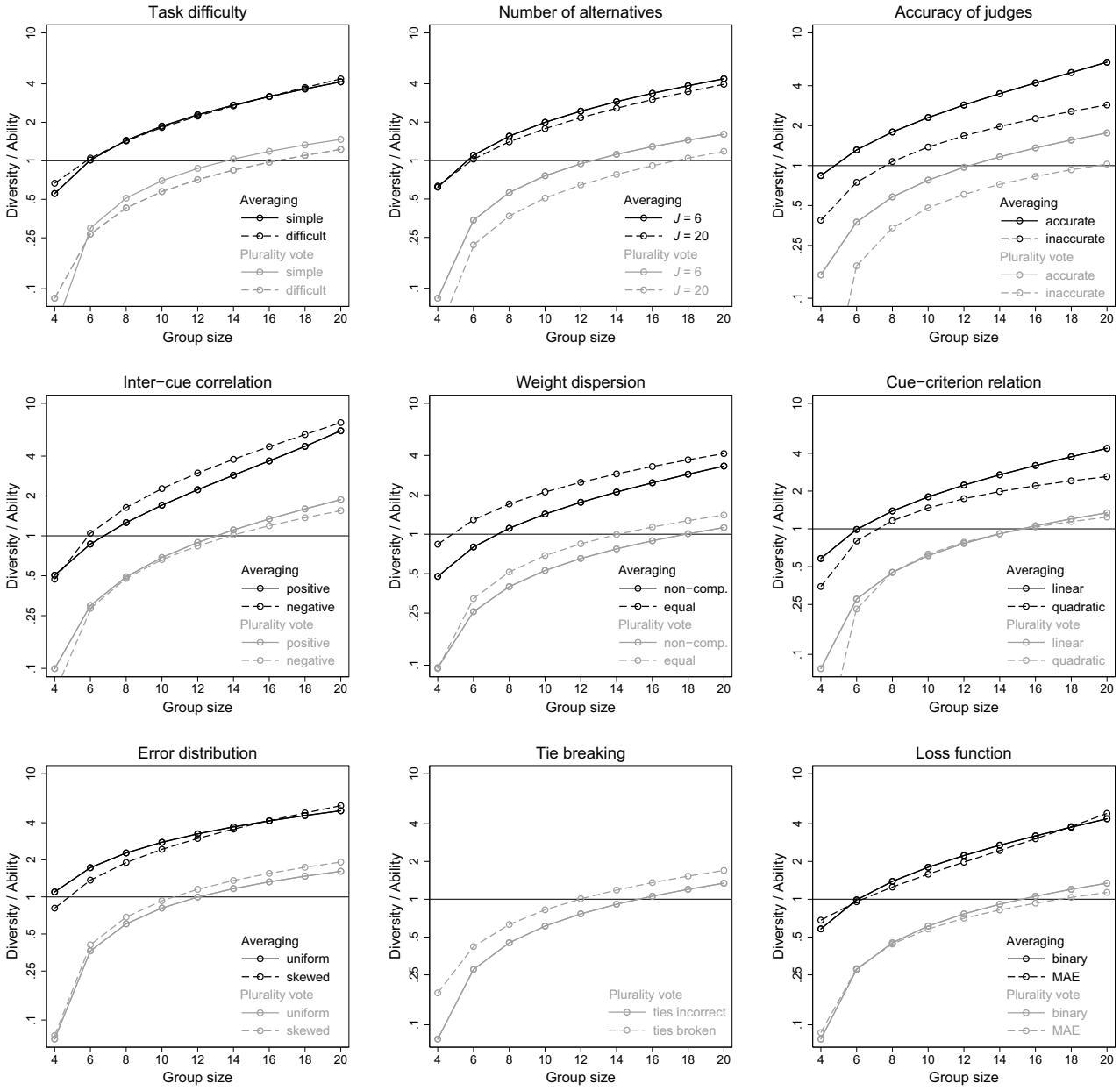
Inter-cue correlation. In our testbed correlations between cues vary randomly. 16.2% (8.7%) of groups decide in an environment in which all cues are positively (negatively) correlated. For these groups, mean correlation of cues is .293 and –.235 respectively. A separate analysis of both environmental types shows opposing effects for both aggregation rules. Still, moderation by inter-cue correlation is limited.

Weight dispersion. We substituted non-compensatory weights $\beta_1 = .7, \beta_2 = .2, \beta_3 = .1$ in equation 1 which, for both aggregation rules, make ability more essential. The special case of equal weights $\beta_1 = .333, \beta_2 = .333, \beta_3 = .333$ yields results similar to our standard compensatory set-up.

Cue-criterion relation. Linear models are applicable to various functional relations in the environment (e.g., Dawes 1979). In most lens model studies in the lab and in the field non-linearities are small and noise is a bigger source of error than failing to capture non-linearities (Karelaia and Hogarth 2008). Still, we set equation 1 to $Q_j = \beta_1 C_{j1}^2 + \beta_2 C_{j2}^2 + \beta_3 C_{j3}^2 + u_j$, employing the special case of quadratic cue-criterion relationships. Effect ratios remain robust, particularly regarding nominal votes.

Error distribution. Our finding is unaffected by the distribution of individual errors. To introduce asymmetric bias in perception e_{ijk} and weighting v_{ik} we drew individual errors from skewed beta distributions (with parameters 2 and 8, skewness .83). To generate random errors at roughly the same intervals as in our standard world we transformed beta distributions to have 0 expected value and to fall into the interval $[-28, 100]$ for e_{ijk} and $[-.7, 2.4]$, $[-.4, 1.4]$, $[-.3, 1.1]$ for v_{ik} respectively. The resulting final errors $\hat{Q}_{ij} - Q_j$ are again (almost) symmetric, which is the reason why results are robust against skewed error distributions. Further, we incorporated a uniform distribution of errors in both e_{ijk} ($[-25, 25]$, 0 expected value) and v_{ik} ($[-.6, .6], [-.4, .4], [-.3, .3]$, each with 0 expected value).

Figure 6 Sensitivity Analysis.



Tie breaking. To create conditions more similar to Hastie and Kameda (2005) we broke tied votes at random. Under plurality rule ties between at least two alternatives occur in 15.3% of groups. Randomization gives 28.1% of tied groups a correct solution. Correspondingly, finding the true alternative becomes less crucial because randomization leads to a substantial proportion of tied votes getting coded as correct in any event. Due to ability's loss in relevance, the effect ratio switches at smaller group size.

Loss function. We substituted a convex loss function for our binary measure of group performance. Following Hastie and Kameda (2005) we use mean absolute error (MAE), i.e. the absolute difference between the best alternative's criterion value and the chosen alternative's value. Our result is remarkably robust against the specific operationalization of group loss.

References

[1] Arrow, K. J. 1951. *Social Choice and Individual Values*. Yale University Press, New Haven.

[5] Balinski, M., R. Laraki. 2010. *Majority Judgment: Measuring, Ranking, and Electing*. MIT Press, Cambridge.

[5] Black, D. 1958. *The Theory of Committees and Elections*. Cambridge University Press, Cambridge.

[5] Bonabeau, E. 2009. Decisions 2.0: The power of collective intelligence. *Sloan Manage. Rev.* **50** 45–52.

[5] Brunswik, E. 1952. Representative design and probabilistic theory in a functional psychology. *Psychol. Rev.* **62** 193–217.

[8] Castellan, N. J. Jr. 1973. Comments on the 'lens model' equation and the analysis of multiple-cue judgments tasks. *Psychometrika* **38** 87–100.

[8] Condorcet, J. A. N. 1785. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Imprimerie Royale, Paris.

[8] Conradt, L., C. List. 2009. Group decisions in humans and animals: a survey. *Phil. Trans. R. Soc. B* **364** 719–742.

[9] Dawes, R. M. 1979. The robust beauty of improper linear models in decision making. *Am. Psychol.* **34** 571–581.

[10] Elster, J. 1989. Taming chance: Randomization in individual and social decisions. J. Elster, ed. *Solomonic Judgments: Studies in the Limitations of Rationality*. Cambridge University Press, Cambridge.

[12] Galton, F. 1907. Vox populi. *Nature* **75** 450–451.

[12] Grofman, B., G. Owen, S. L. Feld. 1983. Thirteen theorems in search of the truth. *Theory Decis.* **15** 261–278.

[17] Hammond, K. R. 1955. Probabilistic functioning and the clinical method. *Psychol. Rev.* **62** 255–262.

[17] Hastie, R., T. Kameda. 2005. The robust beauty of majority rules in group decisions. *Psychol. Rev.* **112** 494–508.

[17] Hogarth, R. M. 1978. A note on aggregating opinions. *Organ. Behav. Hum. Perform.* **21** 40–46.

[17] Hong, L., S. E. Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *P. Natl. Acad. Sci. USA* **101** 16385–16389.

[17] Hong, L., S. E. Page. 2012. Some microfoundations of collective wisdom. H. Landemore, J. Elster, eds. *Collective Wisdom: Principles and Mechanisms*. Cambridge University Press, Cambridge.

[18] Karelaia, N., R. M. Hogarth. 2008. Determinants of linear judgment: A meta-analysis of lens model studies. *Psychol. Bull.* **134** 404–426.

[21] Larrick, R. P., J. B. Soll. 2006. Intuitions about combining opinions: misappreciation of the averaging principle. *Manage. Sci.* **52** 111–127.

[21] List, C., R. E. Goodin. 2001. Epistemic democracy: generalizing the Condorcet jury theorem. *J. Polit. Philos.* **9** 277–306.

[21] Lorenz, J., H. Rauhut, F. Schweitzer, D. Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *P. Natl. Acad. Sci. USA* **108** 9020–9025.

[22] Page, S. E. 2007 *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, Princeton.

[24] Soukhoroukova, A., M. Spann, B. Skiera. 2012. Sourcing, filtering, and evaluating new product ideas: An empirical exploration of the performance of idea markets. *J. Prod. Innovat. Manage.* **29** 100–112.

[24] Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday Books, New York.