Supplementary Information for

An ensemble method utilising multiple thinking styles that boosts the wisdom-of-inner-crowd effect

*Itsuki Fujisaki, Lingxi Yu, Yuki Tsukamura, Kunhao Yang & *Kazuhiro Ueda

*Itsuki Fujisaki & Kazuhiro Ueda
Email: bpmx3ngj@gmail.com (I.F), ueda@g.ecc.u-tokyo.ac.jp (U.K)

This PDF file includes:

S1. Optimal weighting in the Pilot study

Here, we calculated the optimal weighting of the Intuition and Deliberation opinions using the following equation:

$$MSE = (((w \times Intuition + (100 - w) \times Deliberation / 100) - T(true\ answer))^2$$

where *w* represents the weighted percentage of the Intuition. We varied *w* from 0 to 100 in increments of one. Specifically, we set 101 levels (*w* = 0, 1.... .99, and 100) and calculated the MSE for each *w*.

Fig. S1 shows the results. It indicates the relationship between *w* and the MSE. The red dotted line represents the optimal weighting *w* = 47.

The figure also shows a dotted grey line for the average (*w* = 50). It can be observed that these MSEs are almost the same as that in the optimal weighting (0.084% higher). Subsequently, we assumed that the averaging method is a robust strategy.

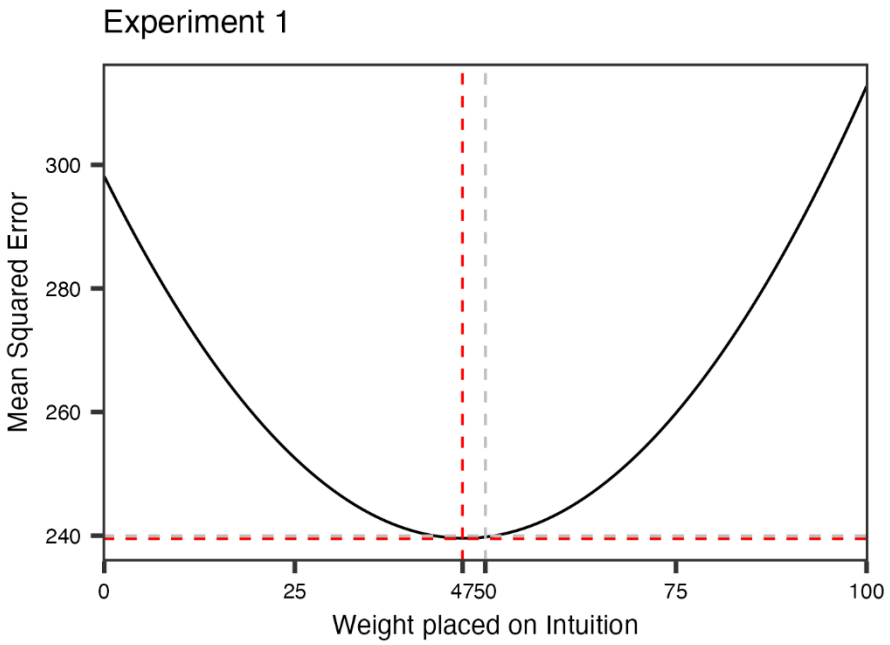


Fig. S1. Results of the optimal weighting in the Pilot study.

S2. Optimal weighting in the Experiment

Here, we focus on the Ensemble method condition in the Experiment. Optimal weighting is expressed as follows:

MSE = (*a* × Intuition + *b* × Deliberation + *c* × Dialectic + *d* × General crowd + *e* × Disagree-other) / 100
a + b + c + d + e = 100

For simplicity, we varied *a, b, c, d, e* from 0 to 100 in steps of 10. Specifically, we set 11 levels (*a, b, c, d, e* = 0, 10.... .90, and 100) and calculated the MSE for each combination of the variables.

These results are summarised in Table S1. It can be observed that the MSE in averaging was almost the same as that in optimal weighting (0.77% higher), as in the Pilot study. Subsequently, we assumed that the averaging method is a robust strategy.

More importantly, for optimal weighting, all variables were above 0. Thus, to exploit accurate estimates, all estimates are needed.

Table. S1. Summary of the analysis.

	a	b	c	d	e	MSE
Optimal weighting	20	30	20	20	10	224.58
Averaging	20	20	20	20	20	226.30

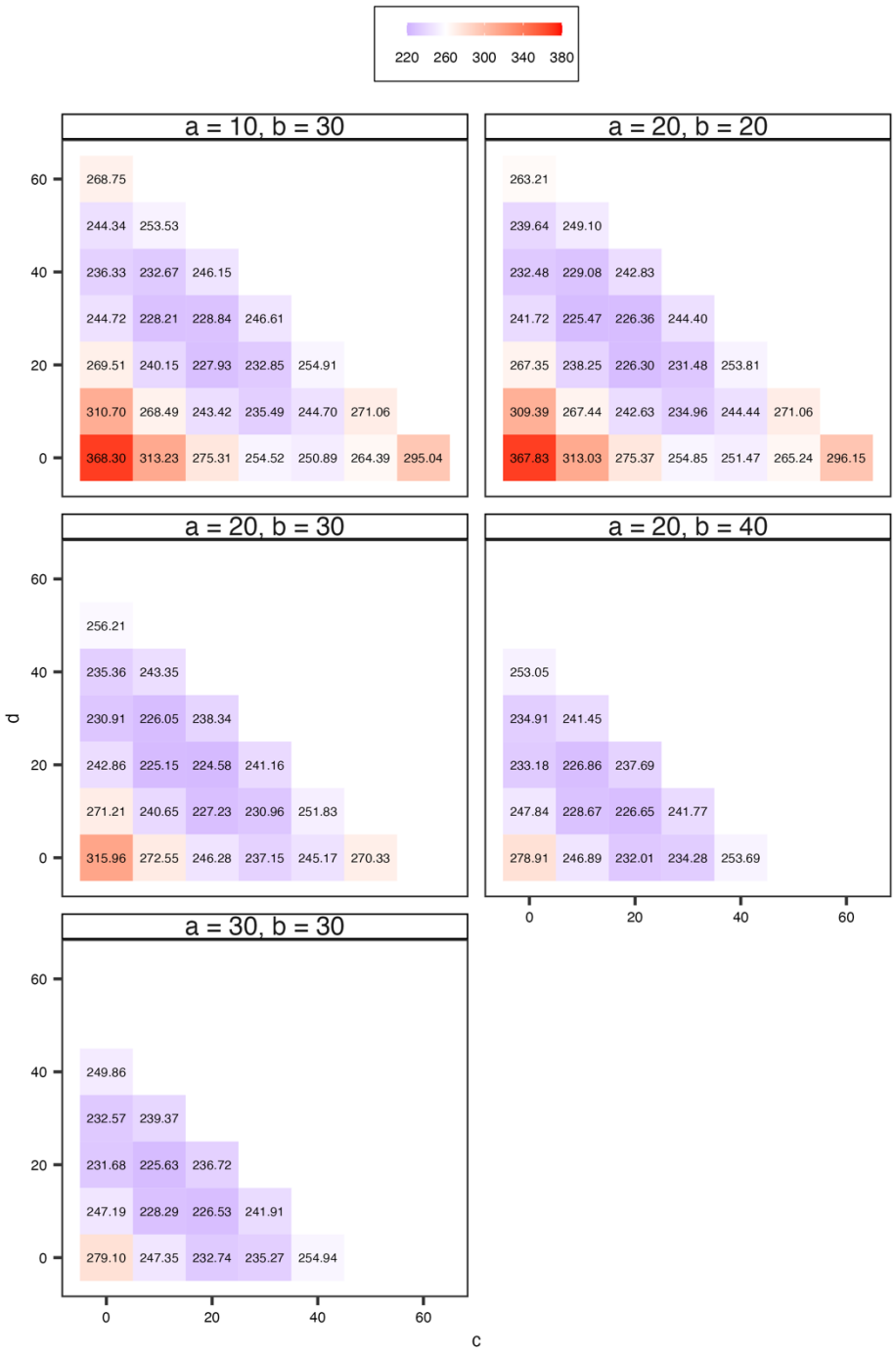


Fig. S2. Results of the optimal weighting in the Experiment. The x axis indicates *c* and the *y* axis represents *d*. Note that *e* is determined by other 4 variables. As shown, the optimal weighting is when *a* = 20, *b* = 30, *c* = 20, *d* = 20, *e* = 10.

S3. Analysis on confidence in the Pilot study

Fujisaki et al. (2023)[1] indicated that confidence in the first estimate influenced the effectiveness of their method. Therefore, we analysed participants' confidence. As mentioned in the Methods section of the main manuscript, participants answered with confidence in the first estimate (i.e. Intuition) between the first and second estimates. Participants rated their confidence on a scale ranging from 0 (I am not confident in my answer at all) to 100 (I am very confident in my answer).

First, we conducted a mixed-effects analysis to examine the influence of confidence. We set $MSE_{average}$ - $MSE_{Intuition}$ (i.e. the effectiveness of this method) as the dependent variable based on Fujisaki et al. (2023)[1], confidence as the independent variable, and the participants and questions as random variables. The results show no significant effect (t = 0.33, p = 0.74).

Additionally, we conducted another mixed-effects analysis with $MSE_{Deliberation}$, confidence, and participants/questions as the dependent, independent, and random variables, respectively. The results showed no significant effect (t = 0.69, p = 0.49).

In short, the results indicated that confidence did not influence the effectiveness of this method.

S4. Brief analysis on MAE (Mean Absolute Error)

Here, as for additional index for accuracy of estimate, we adopted MAE (Mean Absolute Error), instead of. MAE was computed by calculating the absolute value of the difference between the estimate and true value. Simply, we focused on not the squared error but the absolute error.

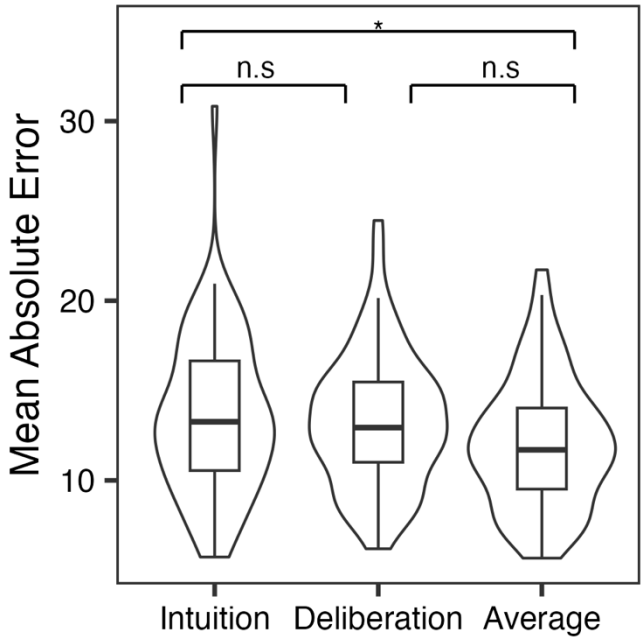
Fig.S3 (Experiment 1) and Table S2 shows the results. We got the same results as the MSE analysis. In the Pilot study, although there was no significant effect between the Deliberation and the Average (t = 0.39; p = .92; *Cohen's d* = 0.35 [0.70, 0.00]), contrary to the MSE analysis, we found the Average estimates were significantly more accurate than that in Intuition (t = 2.72, p = .018; *Cohen's d* = 0.38 [0.73, 0.03]).

In the Experiment, first, we found that the MAE of the average of the five estimates in the Ensemble method condition was lower than that of the first estimate in the Repeated condition (t = 3.64, p < 0.001, *Cohen's d* = 0.53 [1.00, 0.073]). Second, and more importantly, the MAE of the average of the five estimates in the Ensemble method condition was lower than that of the average of the five estimates in the Repeated condition (t = 2.21, p = 0.028, *Cohen's d* = 0.47 [0.93, 0.00]).

We also applied Page's trend test and found the same results as in the main analysis: the MSE in the Ensemble method condition decreased monotonically as the number of estimates increased (z = -5.15, p < .001). By contrast, we confirmed that the MSE in the Repeated condition neither increased nor decreased monotonically (z = 1.09, p = .28).

Pilot study

Fig. S3. Results of the analysis.



Experiment

Table S2. Results of the analysis (MAE). The ranges in brackets represent 95% credible intervals.

Number of guesses	Ensemble method	Repeated
1	15.1 [13.7, 16.7]	13.7 [12.3, 15.1]
2	13.1 [11.8, 14.4]	13.7 [12.3, 15.1]
3	12.6 [11.4, 13.8]	13.8 [12.4, 15.2]
4	12.2 [11.0, 13.5]	13.9 [12.5, 15.3]
5	11.8 [10.5, 13.0]	13.9 [12.5, 15.4]

S5. Analysis on the correlations between estimates in the Experiment

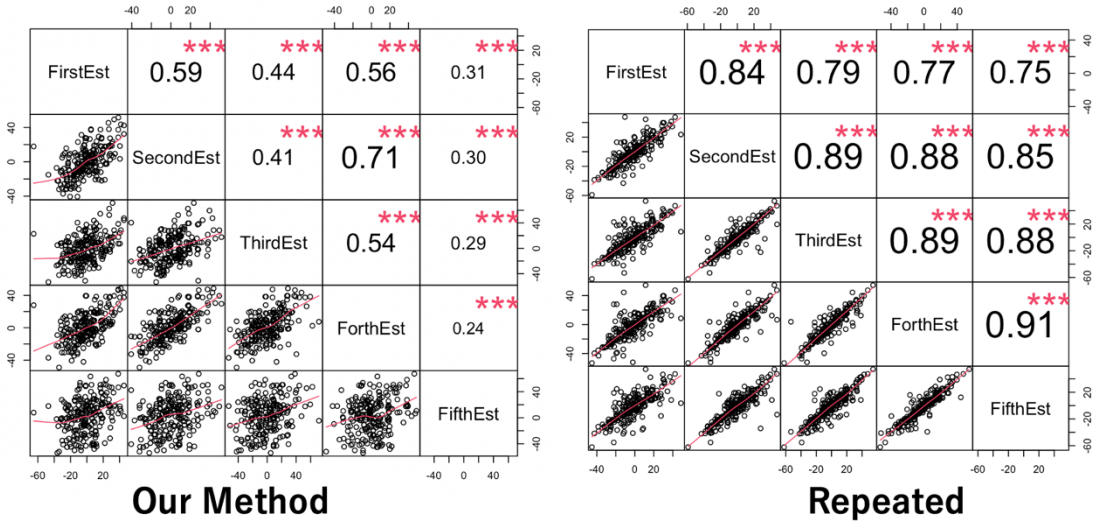


Fig. S4. Results of the analysis. Note that the font size represents the absolute value of correlation coefficient. Our method indicates Ensemble Method condition and Repeated represents Repeated condition, respectively.

S6. Analysis on the accuracy of each estimate in the Experiment

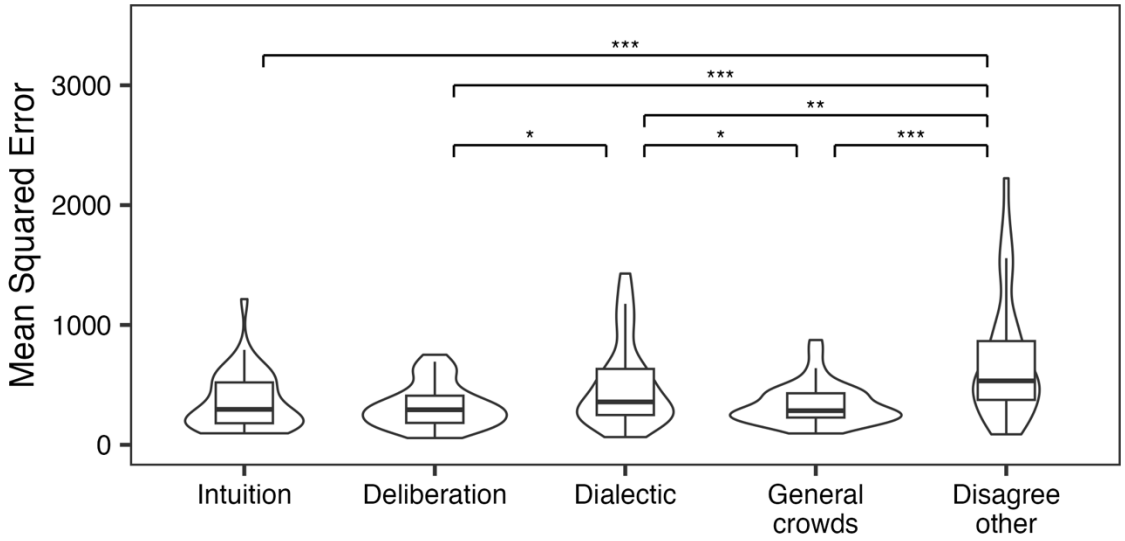


Fig. S5. Results of the analysis on the Ensemble condition. This figure also indicated the results of ANOVA.

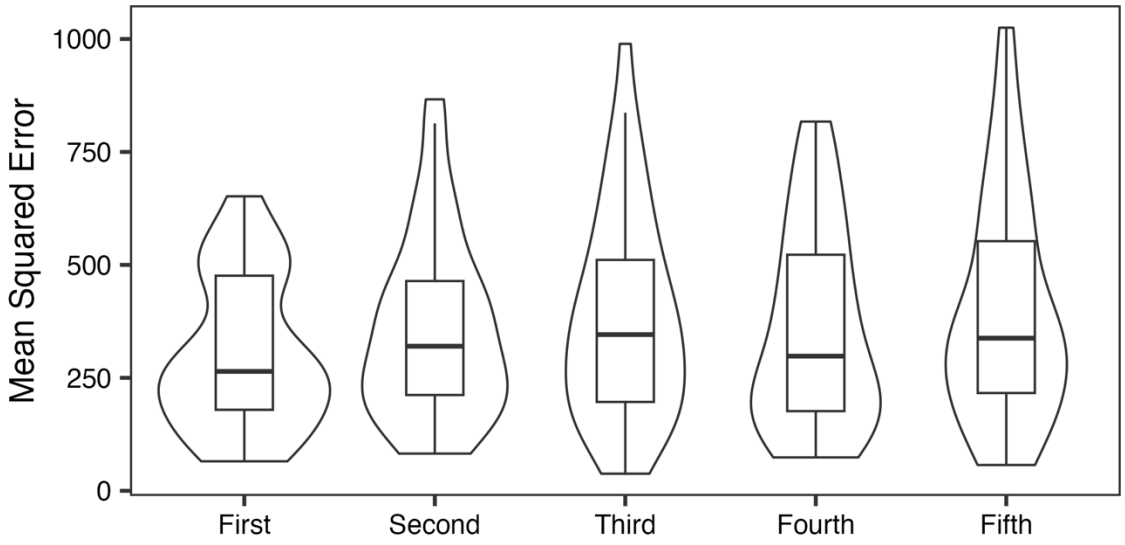


Fig. S6. Results of the analysis on the Repeated condition. This figure also indicated the results of ANOVA (i.e. we found no significant effects).

S7. Analysis on the diversity

Then, we investigated which estimates is most helpful for creating diversity in the ensemble condition. We calculated the *diversity* as follows:

$$Diversity = \left(\gamma - \frac{1}{4} \sum_{\substack{x \in \{Intuition,\ Deliberation,\ Dialectic,\ General\ crowd,\ Disagree\ other\} \\ x \neq \gamma}} x \right)^2$$

γ indicated the value of a particular estimate. As Fig. S7 and Table S3 show, the diversity was high in the estimate by disagree-other instruction, in particular.

Fig. S7 Results of the analysis.

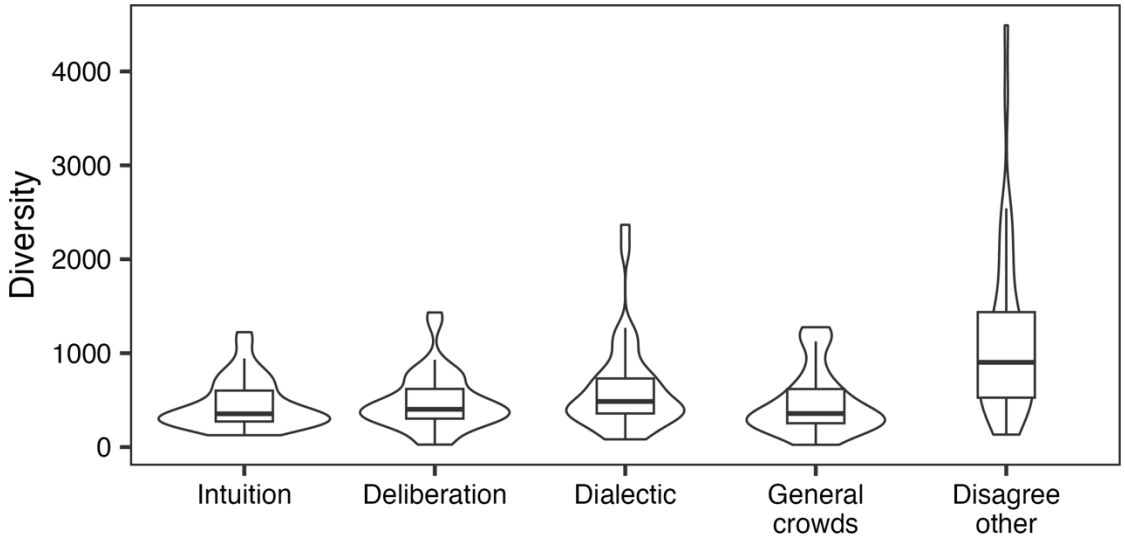


Table S3 Results of the analysis (MSE). The ranges in brackets represent 95% credible intervals.

Estimate	Diversity
Intuition	460.0 [384.5 546.3]
Deliberation	483.1 [399.6 583.4]
Dialectic	623.7 [480.8 798.1]
General crowd's	470.7 [369.6 587.5]
Disagree other	1178.0 [900.1 1505.0]

S8. Additional simulations 1 in the Experiment

In this section, we analysed the influence of the order in which estimates were provided. Specifically, the only difference between the main analysis and this supplementary analysis is that, here, the position of each estimate was randomized across different estimation rounds, consistent with the methodology used in the previous study[2].

In the additional simulations, we randomly selected a person but randomly selecting 1, 2, 3, 4, or 5 estimates (for example, Dialectic, Disagree other, Intuition, Deliberation, General crowds in that order). We repeated the simulations for 1,000 times for each condition.

Fig. S8 shows the results of the simulations. We also display the results in the main text (in Fig.3). We could get similar results to that of the main analysis. When the number of estimates was one, the Simulated Ensemble method condition had larger MSE. However, as the number of estimates increased, the MSE of the Ensemble method condition rapidly decreased. Simulated Repeated condition was relatively flat.

Importantly, the improvement in accuracy from ensemble 1 to ensemble 2 was greater in the Simulated Ensemble condition than in the original Ensemble condition. Likewise, accuracy improved more substantially from ensemble 1 to ensemble 2 in the Simulated Repeated condition compared to the original Repeated condition. This occurred seems to be, in the additional simulations, the 'Disagree-other' estimate—characterized by higher initial errors—could randomly be selected as the first estimate, allowing a larger reduction in error when moving to the second estimate.

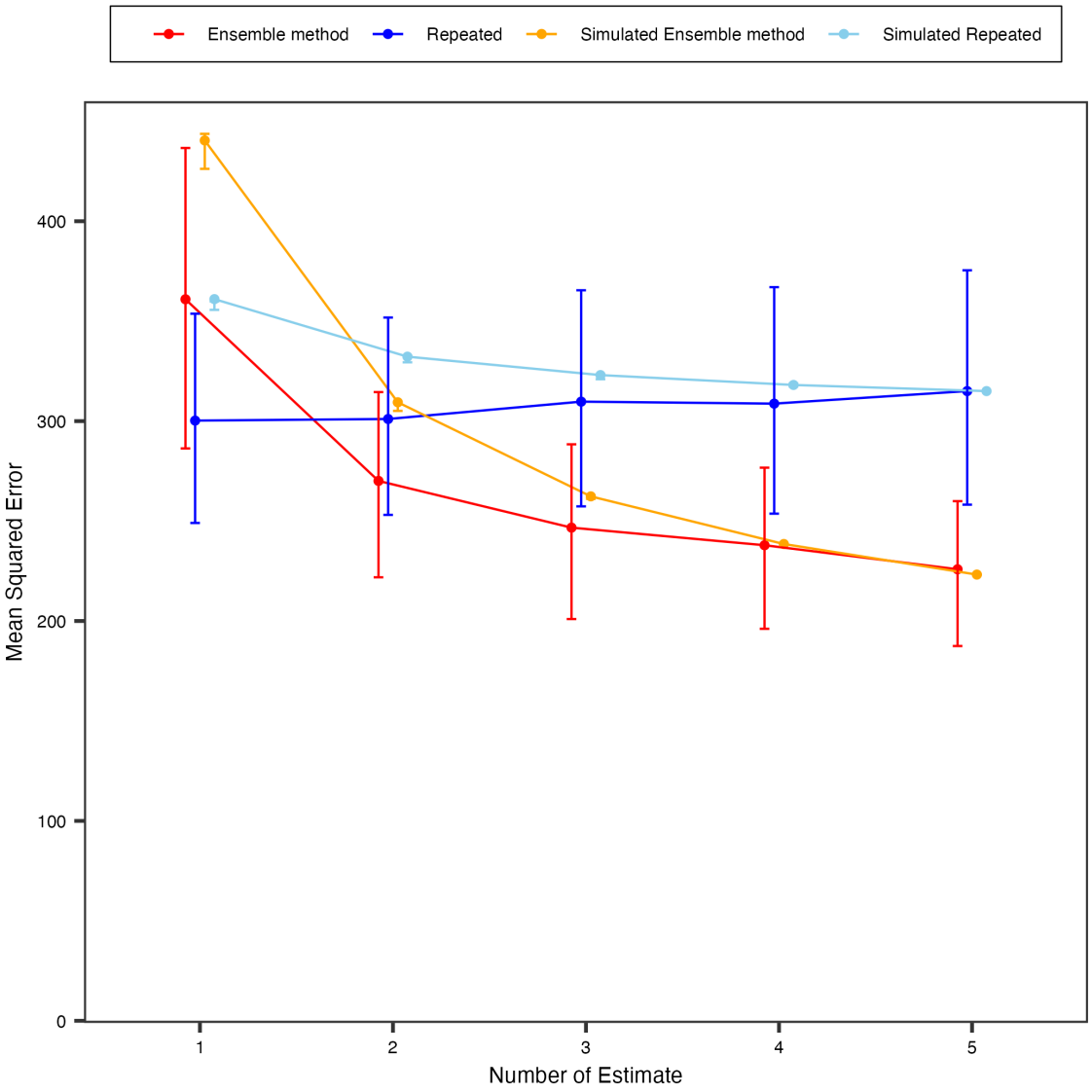


Fig.S8 Results of the additional simulations. We also display the results in the main text (in Fig.3). Bootstrapped 95% confidence intervals (CIs) were computed by resampling the simulated crowd estimates 1,000 times. Note that these CIs can become arbitrarily small if the number of simulations is increased, which is not the statistical generalization target of our analysis.

S9. Additional simulations 2 in the Experiment

In this section, we created the additional outer crowd for the additional benchmark (called 'Ensemble crowd').

In the simulations, we first randomly picked five participants from the ensemble condition. Then, we picked any of estimates of each person (e.g., the second estimate of person A, the fourth estimate of person B etc.), averaged them, and calculated the error. We calculated of the error through error decomposition analysis. We conducted the simulations for 10,000 times.

Table S4 shows the results of simulations. Compared to the Ensemble method, the Ensemble crowd had smaller error in Collective Error, Expected squared error, and Diversity. Especially, we can see that the Ensemble crowd had a much smaller Expected squared error than Ensemble method, which led to the smaller Collective Error. On the other hand, the Ensemble crowd had almost the same Expected squared error as the Repeated.

Table S4. Result of the analysis.

	Collective error	Expected squared error	Diversity
Ensemble method	225.8 [184.9, 270.4]	438.8 [391.7, 486.7]	212.9 [202.8, 223.3]
Repeated	315.0 [259.9, 374.2]	359.0 [303.1, 419.4]	43.9 [39.5, 48.6]
Repeated-first	108.2 [90.2, 127.0]	300.3 [275.4, 325.8]	192.0 [183.8, 200.2]
Ensemble crowd	160.6 [158.8 162.0]	356.9 [350.6 359.6]	233.0 [229.8, 236.1]

S10. The parameter estimation in the Experiment

We first regarded *T* (the number of estimates) as the independent variable and MSE as the dependent variable. We then conducted a parameter estimation (slope *a* and intercept *b*) based on Equation Level 1. Subsequently, we set participant (*s*) and question (*i*) as crossed random variables and then estimated the parameters *a* and *b* (Level 2).

$$\boldsymbol{Level\ 1.}\ \boldsymbol{MSE_{si}} = \frac{\boldsymbol{a}}{\boldsymbol{T_{si}}} + \boldsymbol{b} + \boldsymbol{\varepsilon_{si}}\ \boldsymbol{s} \in \{\mathbf{1}, \ldots, \mathbf{38}\};\ \boldsymbol{i} \in \{\mathbf{1}, \ldots, \mathbf{6}\}$$

$$\boldsymbol{Level\ 2.}\ \boldsymbol{a} = \boldsymbol{\gamma_s^{(a)}} + \boldsymbol{S_s^{(a)}} + \boldsymbol{I_s^{(a)}}$$

$$\boldsymbol{b} = \boldsymbol{\gamma_s^{(b)}} + \boldsymbol{S_s^{(b)}} + \boldsymbol{I_s^{(b)}}$$

Note that Rauhut & Lorenz, (2011)[3] indicated that, theoretically, slope *a* corresponds to diversity, and intercept *b* corresponds to collective performance (MSE). We assumed that the parameters *a* and *b* follow a normal distribution. Subsequently, we set the mean of each prior distribution as the diversity and collective performance when the number of estimates was five. We also set the variance of the prior distributions to 300.

$$\boldsymbol{a} \sim \boldsymbol{Norm}(\boldsymbol{Diversity\_T5}, \mathbf{300})$$

$$\boldsymbol{b} \sim \boldsymbol{Norm}(\boldsymbol{Collective\ performance\_T5}, \mathbf{300})$$

In each estimate, we set the number of chains to four, the number of samplings for each chain to 5,000, the warm-up period to 2,000, and the value of the seed to one. Consequently, we confirmed the convergence of parameters *a* and *b* ($R < 1.1$). We used *brms* package of *R* in parameter estimation.

S8. Categorization on the estimates

Pilot study

Table S5. Results of the analysis.

	Overestimate (%)	Underestimate (%)	Correct (%)	Not estimate (%)
Intuition	53.13%	44.84%	0.47%	1.56%
Deliberation	48.59%	51.25%	0.16%	0.00%

Note: For the 'Not estimate', we excluded from the analysis in the main text.

Experiment

Table S6. Results of the analysis.

	Overestimate (%)	Underestimate (%)	Correct (%)	Not estimate (%)
Intuition	33.33	61.40	0	5.26
Deliberation	29.39	70.61	0	0
Dialectic	39.47	60.53	0	0
General crowd's	38.59	61.40	0	0
Disagree other	44.74	55.26	0	0

Table S7. Results of the analysis.

	Overestimate (%)	Underestimate (%)	Correct (%)	Not estimate (%)
1st estimate	22.37	77.63	0	0
2nd estimate	26.75	72.80	0	0.44
3rd estimate	28.07	71.93	0	0
4th estimate	25.88	74.12	0	0
5th estimate	24.12	75.87	0	0

S11. Estimated values in the Pilot study (Analysis 1)

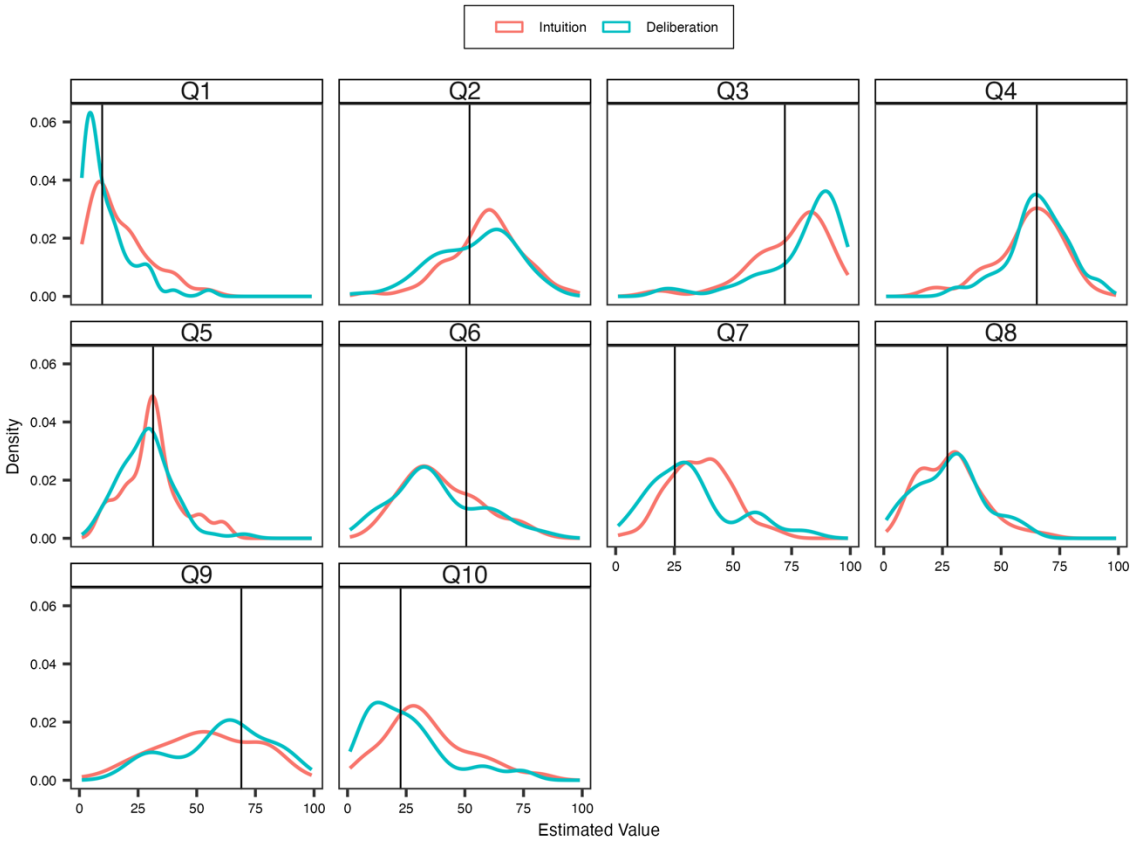


Fig. S9. Result of the estimated values. Each black line indicates the correct answer.

S10. Estimated values in the Pilot study (Analysis 2)

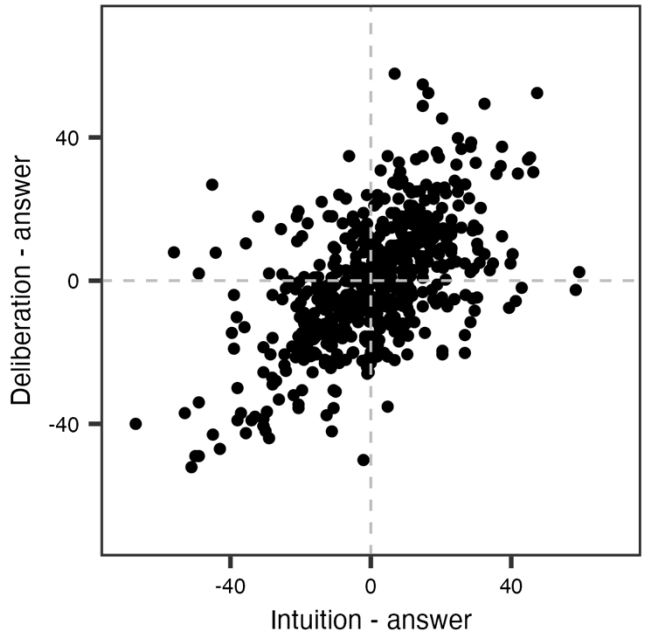


Fig. S10. Result of the analysis. The grey dotted lines indicate that Intuition or Deliberation are equal to the correct answer.

S12. Estimated values in the Experiment

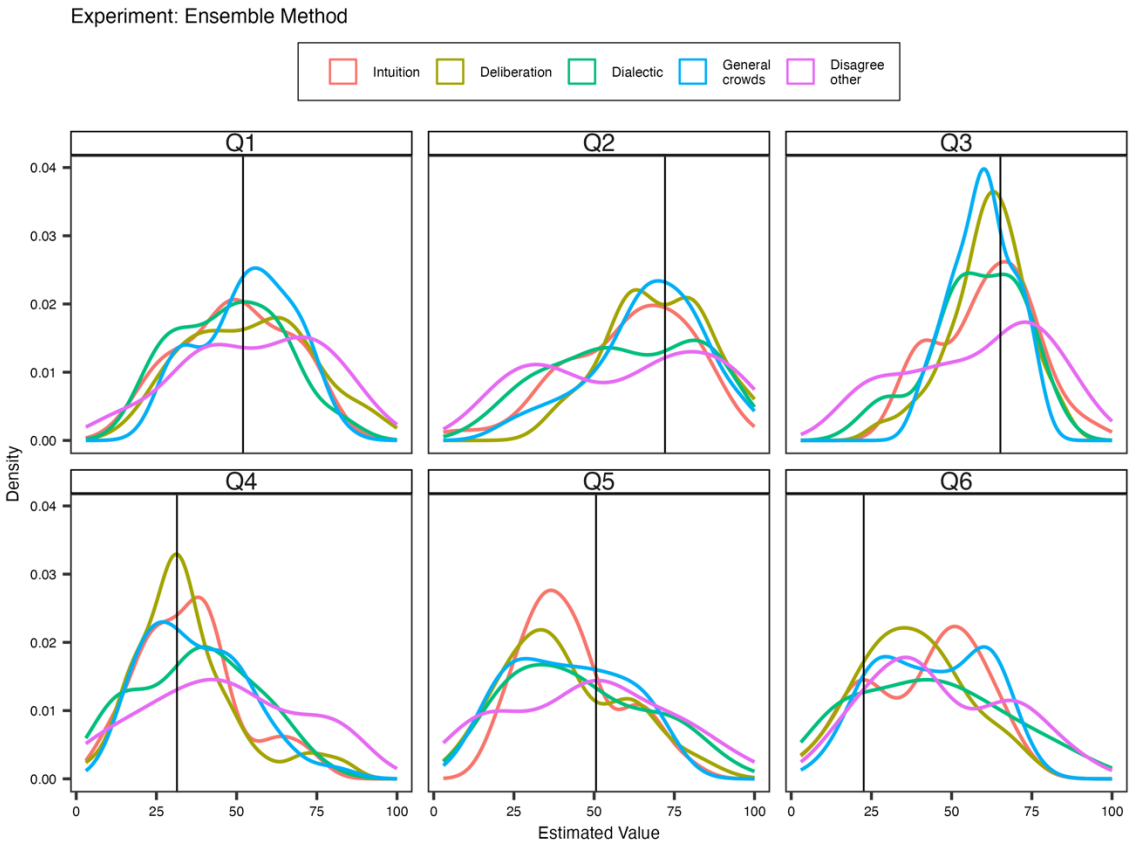


Fig. S11. Result of the estimated values in the Ensemble method condition.

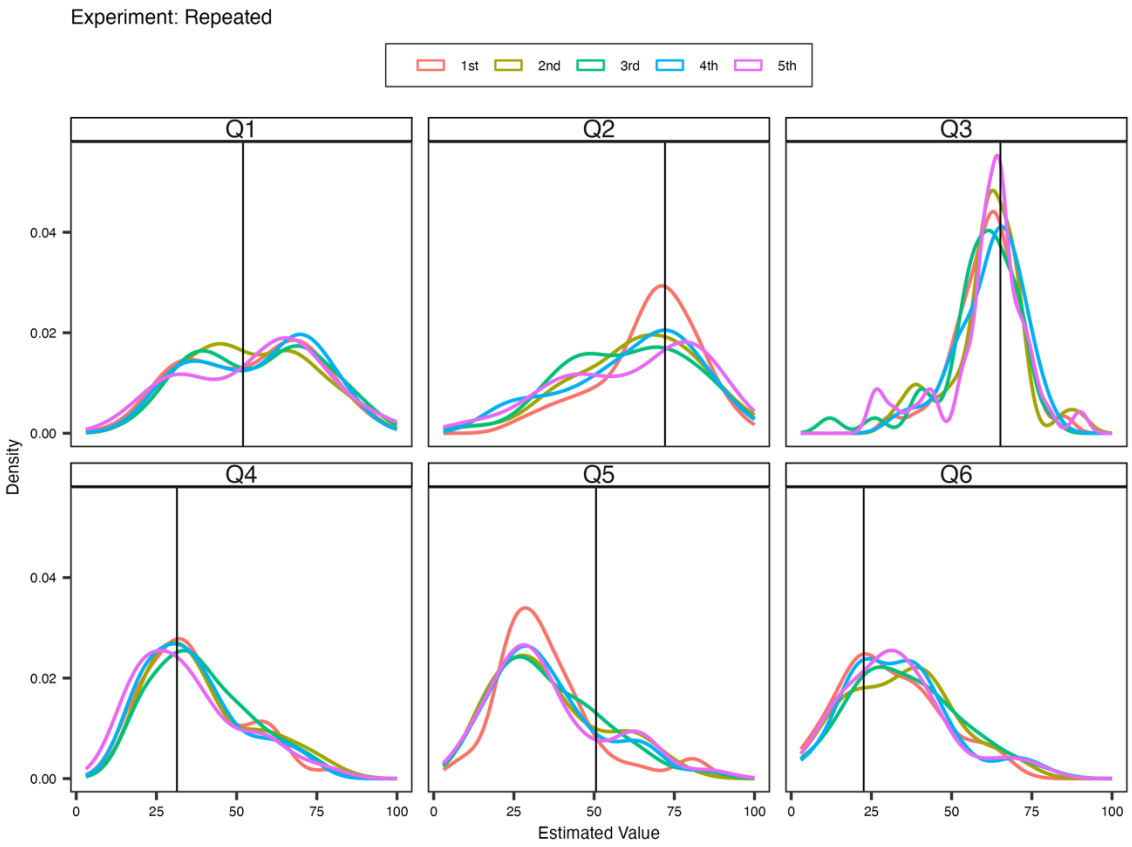


Fig. S12. Result of the estimated values in the Repeated condition.

1. Fujisaki I, Yang K, Ueda K. (2023) On an effective and efficient method for exploiting the wisdom of the inner crowd. *Sci. Rep. 13*:3608. (doi: 10.1038/s41598-023-30599-8)
2. Dolder, D Van., and Assem, M. J. Van Den. (2018). The wisdom of the inner crowd in three large natural experiments. *Nat. Hum. Behav. 2, 21-26.* (doi:10.1038/s41562-017-0247-6)
3. Rauhut H, Lorenz J. (2011) The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *J. Math. Psychol. 55*:191–197. (doi:10.1016/j.jmp.2010.10.002)