

Optimal Group Decision: A Matter of Confidence Calibration *

Sébastien Massoni^{1†} and Nicolas Roux¹

¹ CES - University of Paris 1 and Paris School of Economics

December 2012

PRELIMINARY DRAFT

Abstract

The failure of groups to make optimal decisions is an important topic in human sciences. Recently this issue has been studied in perceptual settings where the problem could be reduced to the question of an optimal integration of multiple signals. The main result of these studies asserts that inefficiencies in group decisions increase with the heterogeneity of its members in terms of performances. In this paper we assume that the ability of agents to appropriately combine their private information depends on how well they evaluate the relative reliability of their information. We run a perceptual experiment with dyadic interaction and confidence elicitation. It gives evidence that predicting the performance of a group is improved by taking into account its members' confidence in their own reliability. Doing so allows us to revisit previous results on the relation between the performance of a group and the heterogeneity of its members' abilities.

Keywords: group decision, metacognition, signal detection theory

*Research for this paper was supported by a grant from Paris School of Economics. The authors are grateful to Steve Fleming, Nicolas Jacquemet, Jean-Marc Tallon, Antoine Terracol and Jean-Christophe Vergnaud for insightful comments.

[†]Corresponding author: sebastien.massoni@gmail.com.

”For difficult problems, it is good to have 10 experts in the same room, but it is far better to have 10 experts in the same head.” John Von Neumann.

Groups are often trusted to make decisions because they gather the information of their members. However, the extent to which groups are able to combine information coming from different sources remain an open question.

Recently, the question has been carried to the field of psychophysics by studying group decision making in *signal detection* experiments (1), (2), (3), and (4). Signal detection experiments consist in asking subjects to make a binary decision based on noisy perceptive information (5). A typical signal detection experiment in groups consists in asking subjects individually and then as a group to tell which one of two visual stimuli was the strongest.

A long standing literature has shown that people’s decisions in this type of situations could be considered as being made by a Bayesian decision maker equipped with some (perceptive) information structure (6) and (7). The modeling of perceptive information makes it possible to determine what would be the performance of a group if it perfectly combined its members’ information (4). Comparing actual group performance to this benchmark, (1) and (2) find that groups whose members are heterogeneous in terms of perceptive abilities (that is one of them has a higher probability of finding the strongest stimulus) tend to perform poorly.

The failure of heterogeneous groups suggests that the reliability of individual information is not well accounted for in the way it is aggregated. (1) propose, as explained by (8), that groups use a suboptimal decision rule that overweights the recommendations of the least able member. The resulting efficiency loss is increasing in the difference in group members’ information reliabilities. This model therefore postulates the existence of a systematic failure in the way private information is aggregated.

On the contrary, we propose to relate those results to biases in subjects’ confidence calibrations. We assume that subjects’ beliefs about their perceptive abilities are initially not related to their actual perceptive abilities. If everyone holds similar beliefs about his performances, the most able subjects tend to be relatively underconfident as compared to the least able subjects (9). Consequently, a group will put too much weight on the least able member, so that heterogeneity induce greater collective inefficiencies. Therefore our explanation of collective inefficiencies does not rely on the incapacity of humans to aggregate heterogeneous information. We rather see them as an inevitable consequence of the lack of information subjects have access to.

Since our aim is to show that inefficiencies are related to subjects’ beliefs about their perceptive abilities, we conduct a signal detection experiment with group deci-

sions in which we elicit subjects' confidence at each trial. A confidence is defined as the subject's belief that he chose the right stimulus. The results support our hypothesis. They are in line with new results that explore the links between metacognitive abilities and group decision (2), (3), and (10).

1 Hypotheses

It is well established in Signal Detection Theory that the perceptive information subjects receive can be fruitfully modeled as a Bayesian information structure (6). A subject draws at each trial a perceptive signal $x \in (-\infty, +\infty)$. Signals are drawn from a normal distribution whose mean, θ , depends on the actual contrast difference between the two stimuli. The variance σ_i^2 captures the precision of subject i 's perception. We will often talk about a subject's precision parameter as the inverse of his variance, $\tau_i = 1/\sigma_i^2$. As we use only one level of difficult, the contrast difference θ can take two values, μ (right stimulus stronger) and $-\mu$ (left stimulus stronger), which are equally likely to occur. Subjects are asked to tell whether θ is positive or negative. Individually, their decision rule is to follow the sign of the signal they receive. The probability that subject i makes the right decision corresponds to the probability that he receives a positive signal conditional on $\theta = \mu$. It is thus given by $\Phi(\mu\sqrt{\tau_i})$ where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

As a group, if subjects perfectly combine their private informations, they make decisions based on the sign of the sum of their signals weighted by the precision of their informations: $x_G = \tau_1 x_1 + \tau_2 x_2$. Note that this statistics is positive if and only if the likelihood of (x_1, x_2) given μ is greater than the likelihood given $-\mu$. The probability of a (optimal) group making a correct choice is thus given by the probability that x_G is positive conditional on μ . x_G is normally distributed with mean $(\tau_1 + \tau_2)\mu$ and precision $1/(\tau_1 + \tau_2 + 2\rho\sqrt{\tau_1\tau_2})$, where ρ is the correlation coefficient between group members' signals.¹ It follows that the ideal group's information precision is given by

$$\tau_G^* = \frac{(\tau_1 + \tau_2)^2}{\tau_1 + \tau_2 + 2\rho\sqrt{\tau_1\tau_2}}$$

According to the findings of (1), the comparison of observed group success rate and its ideal success rate τ_G^* reveals that inefficiencies are positively related to the heterogeneity of the group with respect to the precisions of its members. (1) then propose an alternative model which is based on a suboptimal decision rule (named

¹Other models do not take into account this correlation but as our data show a correlation between signals we integrate it in our model and the alternative ones. See the SI for the computation of the coefficient.

thereafter the suboptimal model). Groups make decisions based on the sign of the statistics $x_G^{sub} = \sqrt{\tau_1}x_1 + \sqrt{\tau_2}x_2$. Weighting each member's signal by the square root of its precision instead of the precision induces group to follow the individual with the lowest precision too often (8). The group precision as a function of its members' precisions is

$$\tau_G^{sub} = \frac{(\sqrt{\tau_1} + \sqrt{\tau_2})^2}{2(1 + \rho)}$$

which corresponds to the optimal case when $\tau_1 = \tau_2$ but gets lower as τ_1 and τ_2 become different, i.e. in case of group heterogeneity.

We propose an alternative model, the belief model, in which the failures of heterogeneous groups comes from a lack of information about their members' precisions. Assume that subject i holds some beliefs about his precision parameter whose expectation is noted $\tau_{i,e}$. We make the approximation that a group decision rule is based on the expected values of precision parameters of its members, i.e. a group chooses right when $\tau_{1,e}x_1 + \tau_{2,e}x_2$ is positive.² In other words, the group behaves as if it were sure that these expected precisions are true. Given that x_i , $i = 1, 2$, is actually distributed with precision τ_i , the group statistics is normally distributed with mean $(\tau_{1,e} + \tau_{2,e})\mu$ and precision $\tau_1\tau_2/(\tau_{1,e}^2\tau_2 + \tau_{2,e}^2\tau_1 + 2\rho\tau_{1,e}\tau_{2,e}\sqrt{\tau_1\tau_2})$. It follows that the precision of such a belief-based group is given by

$$\tau_G^{bel} = \frac{\tau_1\tau_2(\tau_{1,e} + \tau_{2,e})^2}{\tau_{1,e}^2\tau_2 + \tau_{2,e}^2\tau_1 + 2\rho\sqrt{\tau_1\tau_2}\tau_{1,e}\tau_{2,e}}.$$

If subjects' expectations are well calibrated, i.e. $\tau_{i,e} = \tau_i$ for $i = 1, 2$, the belief-based group reaches its optimal precision level, i.e. $\tau_G^{bel} = \tau_G^*$. Actually, subjects may have biased expectations and still reach their optimal collective precision: group decisions are optimal as long as $\tau_{1,e}/\tau_{2,e} = \tau_1/\tau_2$. These expectations could be estimated by eliciting the level of confidence of subjects in their choices. This belief model predicts that the heterogeneity of a group with respect to the precision parameters of its members has no direct impact on the group performance. However, since subjects do not initially know their precision, subjects' expected precisions should be (at least) initially unrelated to actual precisions. This assumption is supported by recent evidence showing that metacognitive ability is dissociable from task performance and varies across individuals (11), (12), and (13). To see this, suppose that for every subject i , $\tau_{i,e}$ is drawn from some distribution that is independent of τ_i . It follows, that whatever the value of τ_1/τ_2 , the expected value of $\tau_{1,e}/\tau_{2,e}$ is 1. As a result, all groups treat their members equally which induce more heterogeneous

²The optimal decision rule is a much more complex object to study as it must take into account the whole beliefs about subjects' precisions. The description of the true rule is given in the SI.

groups to experience greater inefficiencies.

2 Protocol

In order to evaluate our model and test our hypotheses we perform a signal detection task in which we elicit subjects' confidence at each trial (see Figure 1 for the details of the experimental design). This experiment involves a perceptual task of numerosity: subjects observe during a short time interval two visual stimuli and are asked to tell which stimulus was stronger and the level of confidence they have in this decision. Each subject of a dyad answers individually, and then the two group members must reach an agreement (on choice and confidence). After the group decision is made, subjects answer anew to check if they agree with group decision. Finally they observe whether they were right or wrong. This sequence, that we will refer to as a trial, is repeated 150 times.

3 Data Treatment

We present results based on 33 groups.³ We have presented the models in terms of precision parameters. We will present our results using directly success rates, s , which is equivalent since the precision parameter completely determines success rate. Indeed our experiment features only one level of contrast difference between the two stimuli so that a subject's success rate fully characterizes his perceptive information precision.

We start by checking the assumption that subjects' expected success rates are not related to their actual success rates (i.e. τ_e is not related to τ). Subject i 's expected success rate, noted $s_{i,e}$, is assumed to be equal to the mean of his reported confidences (recall that we fix the stimulus and thus that confidence is only driven by precision). By regressing the actual success rates s on s_e we do not obtain any correlation between these two variables (see Figure 2A). As expected this implies that subjects that perform well tend to be relatively underconfident as compared to those performing poorly. The linear regression of individual miscalibration, defined as $s_{i,e} - s_i$, on individual success rate indeed shows that those two variables are significantly positively related (see Figure 2B).

Then we examine whether the main result of previous experiments, namely that heterogeneity in group members' success rates impairs group performance, holds in our experiment. Based on the observed success rates of the group members', s_1 and

³Methods, descriptive statistics and additional results supporting our conclusion can be found in the SI.

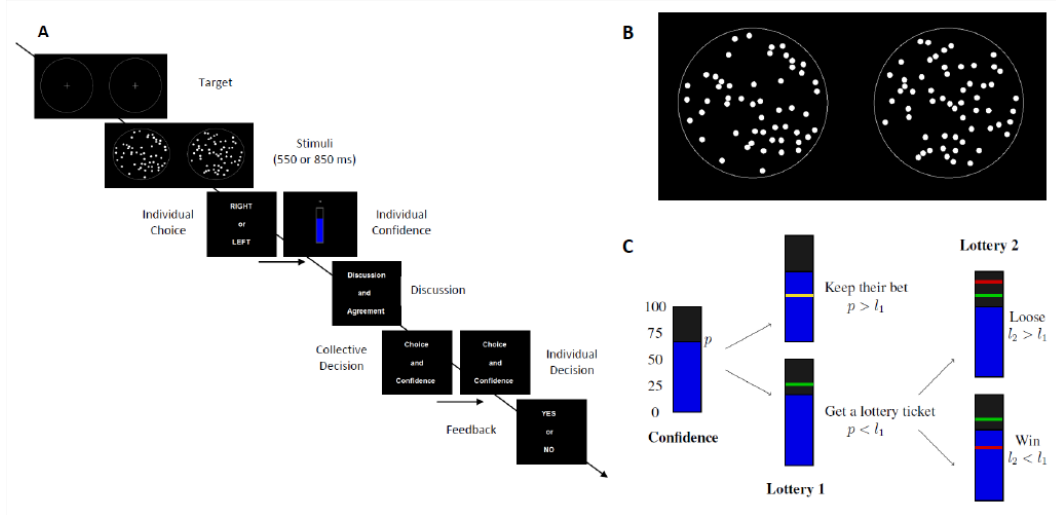


Figure 1: Design of the experiment. **(A)** Experimental paradigm. The experiment is organized as follow: subjects perform a numerosity task in which they observe two circles containing a certain number of dots during a short interval of time, so that it is impossible to count the dots **(B)**. They first tell which one of the two circles is the most likely to contain more dots. Since we are interested in measuring the source of collective loss due to confidence calibration, we elicit confidences at each trial through a matching probability rule **(C)**. The basic principle of this elicitation rule is to elicit an objective probability equivalent to a subjective one. Specifically, confidence takes values between 0 and 100 (by steps of 5) and represents subjects' estimated probability of success at each trial. Note that the reported confidence has a cardinal value. Thus we assume that subjects could report real probabilities as subjective beliefs and not only some ordinal values of feeling. That is why a subject's remuneration during the experiment depends upon the consistency of reported confidence with actual performance. After each trial, subjects observe whether or not they chose the right circle. Subjects make a sequence of 50 trials in isolation. In order to guaranty enough heterogeneity in the groups, half of subjects observe the circle during a shorter time interval than the other half of subjects. Groups of two subjects (with different observation times) are then formed and make 150 trials again. For each trial, subjects independently observe the same two circles and make individual decisions (choice and confidence). The group members are then asked to reach an agreement on each of the two decisions by free communication. After the group decisions are made, each group member reports individual decisions again so that we can check whether group members agreed with the group decisions.

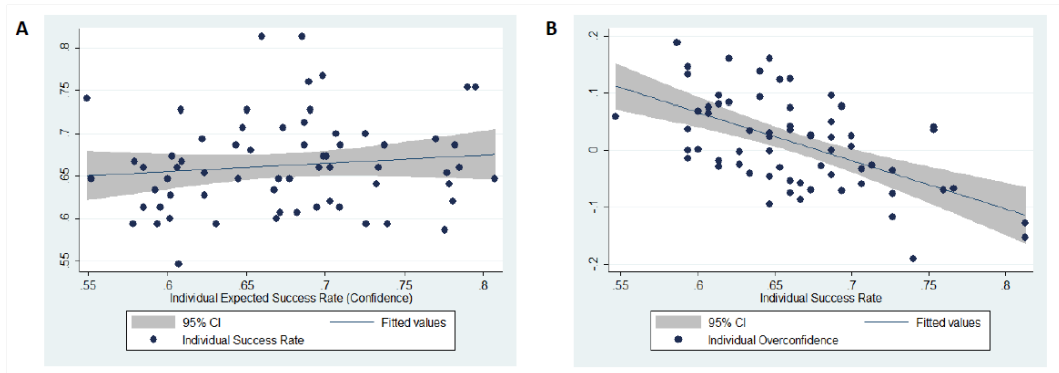


Figure 2: Tests of the main assumptions. **(A)** No relation between expected success rate and actual success rate. We check our first assumption that confidence is not related to real success rate. The graph represents the link between individual success rate and individual confidence. We do not obtain any relation between these two variables. An OLS regression of confidence on success provides a slope of 0.09 with a p-value of 0.302 ($n = 66$). **(B)** Positive relation between perceptive ability and miscalibration. As we have proved that confidence is disconnected to success rate, it follows that subjects with a higher precision parameter tend to be more underconfident than subjects with lower abilities. The graph represents the relationship individual success rate and individual miscalibration. This link is confirmed by an OLS regression of success rate on miscalibration: the coefficient takes a value of -0.85 with a p-value of 0.000 ($n = 66$).

s_2 , and on the estimated correlation coefficient $\hat{\rho}$, we compute the optimal group success rate, s_G^* . We define *collective losses* as the difference between s_G^* and the actual group success rate s_G . Heterogeneity in members' precisions is defined as the absolute value of the difference between members' success rates: $|s_1 - s_2|$. An OLS regression of collective losses on group heterogeneity provides a positive coefficient of 0.32 that is statistically significant ($t = 1.75$, p-value = 0.089 - cf. the regression 1 in Table 1).

Table 1: Impact of group heterogeneity and belief-based losses on collective losses.

Collective Losses ($n=33$)	Regression 1	Regression 2
Group Heterogeneity	0.32**	0.13
Difference in Miscalibration	.	0.25**
Constant	-0.01	-0.01

We now present evidence that the relation between heterogeneity of a group and its collective losses runs through belief miscalibration. Regressing the collective losses on the difference of miscalibration between members and group heterogeneity provides the following results: an effect of miscalibration statistically significant (coefficient of 0.25 with $t = 2.18$, p-value = 0.037) and a removal of the previous link between group heterogeneity and collective losses (cf. the regression 2 in Table 1). Therefore, the relation between group heterogeneity and collective losses disappears when we control for miscalibration. We conclude that heterogeneity in group members' information precisions only impairs group performance if beliefs are miscalibrated.

We now test our model against the optimal model and the suboptimal model. All three models make a prediction about the group success rate so we can compare the explanatory power of the three models on the observed success rate s_G . We perform separate OLS regressions of s_G on s_G^* , s_G^{sub} and s_G^{bel} (regressions (a), (b) and (c), respectively, in Table 2).

Table 2: OLS regressions of the actual group success rate on each model's predictions.

Actual Success s_G ($n=33$)	(a)	(b)	(c)
Optimal Model s_G^*	0.83***	.	.
Suboptimal Model s_G^{sub}	.	0.51***	.
Belief Model s_G^{bel}	.	.	0.94***
Constant	0.11	0.36***	0.05

We compare the resulting R^2 : the belief model provides a $R_{bel}^2 = 0.6266$ while the suboptimal model and the optimal model yield $R_{sub}^2 = 0.4745$ and $R_*^2 = 0.5610$

respectively. We perform a Vuong test of R^2 (14) and our model has a statistically significant higher explanatory power than the suboptimal one (Vuong z-statistic = -2.3033 , p-value = 0.0213) and closed to significant against the optimal one (Vuong z-statistic = -1.4204 , p-value = 0.1555).

4 Conclusion

We propose a model of (approximately) optimal group decision incorporating group members' miscalibration. We are only interested in problems of miscalibration but we can expect that discrimination abilities play a role in group decisions. Discrimination refers to the ability of an agent to distinguish between two signals of different values. Limited discrimination abilities suggest that perceptive signals are filtered before being accessible to the individual so that the assumption that signals are perfectly observed should be weakened. An optimal model of group decision based on confidence should incorporate these two aspects. (3) proposes a model in which group decision is lead by the more confident member. This type II optimal model leads on our data to an underestimation of the group performances. It would therefore be worth investigating the relation between group members' metacognitive abilities (calibration and discrimination) and group performance.

References

- [1] Bahrami et al. (2010) Optimally Interacting Minds. *Science* 329(5995): 1081-1085.
- [2] Bahrami et al. (2012) What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1594): 1350-1365.
- [3] Koriat A (2012) When Are Two Heads Better than One and Why?. *Science* 336(6079): 360-362.
- [4] Sorkin R D, Hays C J, West R (2001) Signal detection analysis of group decision making. *Psychological Review* 108: 183-203.
- [5] Faisal A A, Selen L P J, Wolpert D M (2008) Noise in the nervous system. *Nature Reviews Neuroscience* 9: 292-303.
- [6] Green D A, Swets J A (1966) Signal Detection Theory and Psychophysics. *Springer Series in Statistics, John Wiley and Sons*.

- [7] Ma W J, Beck J M, Latham P E, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9: 1432-1438.
- [8] Ernst M O (2010) Decisions Made Better. *Science* 329(5995): 1022-1023.
- [9] Kruger J, Dunning D (1999) Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6): 1121-1134.
- [10] Frith C D (2012) The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1599): 2213-2223.
- [11] Fleming S M, Weil R S, Nagy Z, Dolan R J, Rees G (2010) Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science* 329(5998): 1541-1543.
- [12] Rounis E, Maniscalco B, Rothwell J C, Passingham R E, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience* 9(8): 165-175.
- [13] Fleming S M, Dolan R J (2012) The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1594): 1338-1349.
- [14] Vuong Q H (1989) Likelihood Ratio Tests for Model Selection and Non Nested Hypotheses. *Econometrica* 57(2): 307-333.

Supplementary Information

Method

Participants This experiment was conducted in May and June 2012 at the Laboratory of Experimental Economics in Paris (LEEP) of the University of Paris 1. Participants were recruited by standard procedure in the LEEP’s database. 35 dyads i.e. 70 subjects (most of them were undergraduate students from University of Paris 1) participated in the experiment for pay. We have lost the data of two groups due to a problem with a computer during the experiment. As we choose to exclude any outliers the data analysis is based on 33 groups of 2 subjects. The experiment last around 2 hours and subjects were paid on average 17 euros.

Materials This computer-based experiment uses Matlab with the Psychophysics Toolbox version 3 (1) and has been achieved on computers with 1024x768 screens.

Task, stimuli and procedure The experimental design is summarized by the Figure 1. The perceptual task is a two-alternative forced choice (2AFC) which is known to be a convenient paradigm for SDT analysis (2). In our task, subjects have to compare the number of dots contained in two circles. We use such numerosity task in reason of neuronal evidence that the brain performs as assumed by SDT facing numerosity problems (see (3) for a review). The two circles are only displayed for a short fraction of time, about one second, so that it is not possible to count the dots. Subjects have to tell which circle contains the higher number of dots and then, their confidence in the choice made is elicited. The time line of the experiment is the following: participants are randomly assigned by dyad, they first make 50 trials in isolation reporting their choice and confidence. Then they face the same stimuli in each dyad for 150 trials. To ensure enough heterogeneity in each group we add a difference in the time presentation of the stimuli: one member had the stimuli during 850 ms while the other member had it during 550 ms. After observing the stimuli they give individually their choice and their level of confidence, then they have to discuss with the member of the dyad to reach an agreement (they have to reveal their choice and confidence but they are free to discuss as long as they want with their partner), they report group’s choice and confidence, and finally they give again their individual choice and confidence in order to check if they agree with the group choice. At the end of each trial, feedback on accuracy is given.

To answer to the choice they have to press one key (F or J) depending on the answer (left or right) and then give their confidence on a scale with value between 0 and 100 with steps of 5. Subjects are paid according to the accuracy of the stated

confidence. We use a probability matching rule (Figure 2) (see (4) and references therein for its properties and implementation) and subjects accumulate some points at each trial (+1 for a correct answer and -1 for a wrong one). The final payment comprises 5 euros of flat payment, the total number of points accumulated during the trials in isolation more the total of points accumulated during the group period (we randomly chose one of the three decisions). All points are converted to the exchange rate: 1 point equal 10 cents.

Model and Analysis

Modelization The coefficient of correlation between the members stimuli that we add in the three models is computed as follows: we observe the probability of the two group members to make the right decision simultaneously. According to the model, this probability should be equal to the probability of both individual signals being positive given $\mu = 1$. Conditional on $\mu = 1$, the distribution of the pair of signal is a bivariate normal with mean $(1, 1)$ and covariance matrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. ρ takes the value that equalizes the theoretical and observed probabilities of group members being simultaneously right. Note that Sorokin et al. (5) propose also to take into account the correlation into their modelization.

In our model we make the approximation that a group decision rule is based on the expected values of precision parameters of its members, i.e. a group chooses right when $\tau_{1,e}x_1 + \tau_{2,e}x_2$ is positive. The exact optimal decision rule must take into account the whole beliefs about subjects' precisions. Noting subject's beliefs about τ_i by Γ_i , the optimal decision rule depends on whether the group posterior about $\theta = \mu$

$$P(x_1, x_2) = \int_0^\infty \int_0^\infty P(x_1, x_2; \tau_1, \tau_2) d\Gamma(\tau_1; \alpha_1, \beta_1) d\Gamma(\tau_2; \alpha_2, \beta_2)$$

is higher or lower than .5.

Learning Effect The results presented in the paper show that if subjects perfectly knew their relative abilities, collective inefficiencies would be statistically independent of collective inefficiencies. The reason why we observe a relation is that subjects have initially no information about their abilities in the task so that on average well performing subjects tend to be relatively underconfident as compared to poorly performing subjects. As subjects repeat the task, they should be able to learn about their ability. The pace at which learning occurs depends upon the feedbacks they receive, but eventually they should be able to combine informations of different reliabilities. In our experiment, subjects observe whether they made the right choice

after each trial. But we will show that the 150 trials were not enough to observe significant improvement in calibration.

We compute each subject’s expected success rate over the first (period 1) and last 75 trials (period 2). Let us note these two expected success rates s_e^1 and s_e^2 respectively. We compute subjects’ success rates s_1 and s_2 over the same periods. Subjects’ expected success rate is not closer on average to their actual success rate: the average miscalibration in period 1, $|s_e^1 - s^1|$ is 0.0679 while it is 0.0752 in second period. A t-test of difference shows that this difference is not statistically significant ($t = -0.8216$, p-value = 0.2072).

As a result, the fact that well performing agents are relatively underconfident remains true throughout the experiment. Table 1 presents the results of the regressions of subjects’ miscalibration in period 1 and 2 over their actual success rate in that period. The relation is significantly negative in both cases.

Actual Success ($n=66$)	Period 1	Period 2
Miscalibration	-0.44***	-0.48***
Constant	0.66***	0.68***

Table 1: Relations between miscalibration and actual success rate in periods 1 and 2.

Therefore we find no evidence of trends in subjects’ calibration. This is the reason why the analysis of this paper is performed using a single calibration estimation for each individual.

Data analysis The Table 2 summarizes the main statistics about the different decisions in the experiment.

	Decision in Isolation	Individual Decision 1	Collective Decision	Individual Decision 2
Success Rate	65.5%	66.3%	69.9%	71.3%
Confidence	70.5%	66.7%	71.1%	73.1%
Calibration	+5.2%	+1.4%	+1.2%	+1.9%
ROC	0.581	0.597	0.649	0.653
Agreement in Choice	.	62.6%	.	85.7%
Agreement in Confidence	.	19.5%	.	33.9%

Table 2: Mean levels of accuracy, confidence, metacognitive abilities and rate of agreement during the different stages of the experiment.

The Figure 3 shows the mean values of the actual success rate of groups compared to the predictions of the different models. The optimal, the suboptimal and

the optimal belief (i.e. choice is lead by the more confident member) models have statistically significant differences with the actual rate. Only our belief model has a non significant difference with actual success.

As Barhrami et al. (6) present their model without taking into account the correlation between group members signals, we check if our main result is robust to analysis with a coefficient ρ equal to zero. Table 3 presents the OLS regressions of group heterogeneity and belief-based collective losses on collective losses. We find the same pattern of results as observed with correlation in the paper (cf. Table 1 of the paper).

Collective Losses ($n=33$)	Regression 1bis	Regression 2bis
Group Heterogeneity	0.31*	0.10
Difference in Miscalibration	.	0.28**
Constant	0.00	0.01

Table 3: Impact of group heterogeneity and belief-based losses on collective losses without correlation.

To test the explanatory power of our model against the optimal and the suboptimal ones we can perform a test of correlations (7) in addition to the test of R^2 . The correlations between the actual success group s and the three predictions s_G^* , s_G^{sub} and s_G^{bel} give support to our model. This correlation is statistically significant higher for our model than the suboptimal one ($z = -1.7596$, p-value = 0.0392) and closed to significantly higher than the optimal one ($z = -0.8747$, p-value = 0.1909).

In order to see that our belief model captures the impact of miscalibration on collective losses, we can run similar regressions to Table 1 of the paper with inefficiencies explained by our model rather than difference of miscalibration. We compute the belief-based success rate of each group, noted s_G^{bel} . This predicted success rate allows us to make predictions on the amount of collective losses a group should exhibit due to the biases of its members' beliefs. Let us call $s_G^* - s_G^{bel}$ the *belief-based collective losses*. Regressing the collective losses on the belief-based collective losses and group heterogeneity provides the following results: an effect of belief-based collective losses statistically significant (coefficient of 0.64 with $t = 2.75$, p-value = 0.010) and a removal of the previous link between group heterogeneity and collective losses (cf. Table 4). Therefore, our model controls for miscalibration and confirms the result of the paper (namely that heterogeneity in group members' information precisions only impairs group performance if beliefs are miscalibrated).

Collective Losses ($n=33$)	Regression 1	Regression 2
Group Heterogeneity	0.32**	0.08
Belief-Based Collective Losses	.	0.64***
Constant	-0.01	0.01

Table 4: Impact of group heterogeneity and belief-based losses on collective losses.

References

1. Brainard D H (1997) The Psychophysics Toolbox. *Spatial Vision* 10: 433-436.
2. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen J D (2006) The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review* 113(4): 700-765.
3. Nieder A, Dehaene S (2009) Representation of Number in the Brain. *Annual Review of Neuroscience* 32 (1): 185-208.
4. Gajdos T, Massoni S, Vergnaud J-C (2012) Elicitation of Subjective Probabilities in the Light of Signal Detection Theory. *Working Paper*.
5. Sorkin R D, Hays C J, West R (2001) Signal detection analysis of group decision making. *Psychological Review* 108: 183-203.
6. Bahrami et al. (2010) Optimally Interacting Minds. *Science* 329(5995): 1081-1085.
7. Cohen J, Cohen D (1983) Applied Multiple Regression / correlation analysis for the behavioral sciences. *Hillsdale JJ: Erlbaum*.

Figures

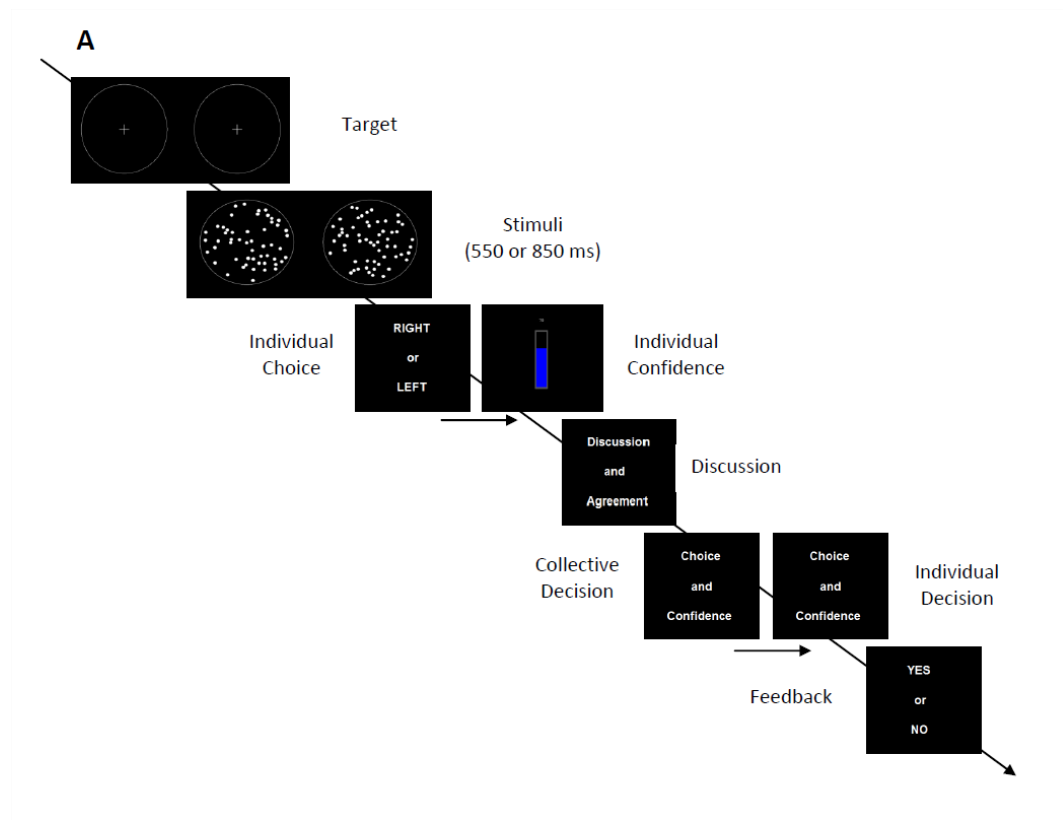


Fig. S1. Experimental design. A group trial in the experiment is defined as follow: subjects see a target and press a key when they are ready; a stimuli of a numerosity task appears during 550 or 850 ms; they make individual decision with choice (right or left) and confidence (on a gauge between 0 and 100); a screen resumes their decision and they have to find an agreement with their partner by free communication; they enter the group decision (choice and confidence) and after that give anew their individual decision (choice and confidence); finally feedback on the accuracy of their answer is provided. Subjects perform overall 150 trials defined as above. In addition they first do 50 trials of only individual decisions.

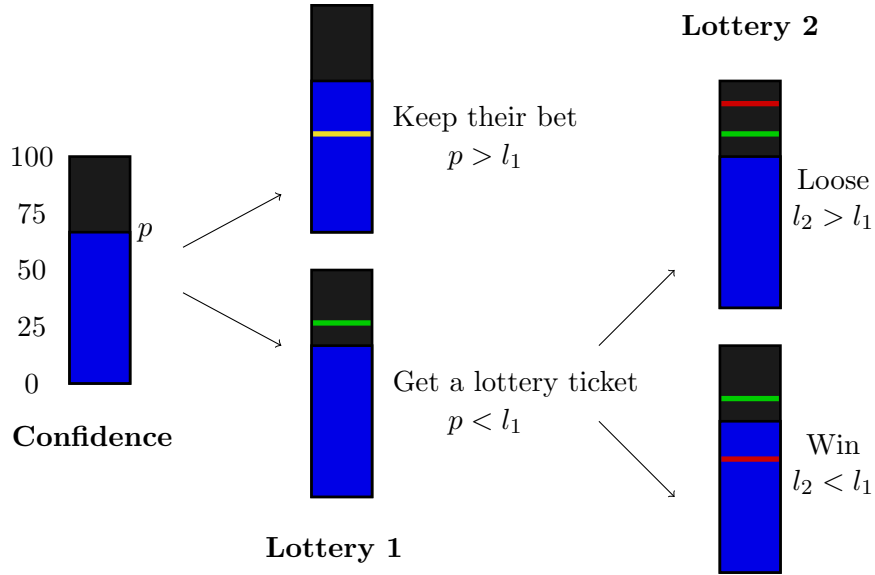


Fig. S2. Confidence elicitation mechanism using probability matching. The principle is to elicit an objective probability equivalent to a subjective one. In our design, subjects have to report on a gauge the probability p that makes them indifferent between a lottery which gives a positive reward in case of a correct answer and a lottery with a probability p of winning the same reward. After the subject has reported a probability p , a random number q is drawn. If q is smaller than p , the subject keeps his initial lottery based on his answer, if q is greater than p , the subject is paid according to a lottery that provides the same reward with probability q . In practice this scoring rule is implemented using a 0 to 100 scale, with steps of 5. Subjects are told that an answer make them hold a lottery ticket based on their answers' accuracy : it gives 1 point if the answer is correct and -1 otherwise. Then on the 0 to 100 gauge, subjects have to report the minimal percentage of chance p they require to accept an exchange between their lottery ticket and a lottery ticket that gives p chance of winning. A number l_1 is determined using a uniform distribution between 0 and 100. If l_1 is smaller than p , subjects keep their initial lottery ticket, if l_1 is higher than p , then the exchange is made with a lottery ticket which gives l_1 chance of winning. In this case, a random draw determines the payment: a number l_2 is determined using a uniform distribution between 0 and 100, the lottery is winning if l_1 is higher than l_2 . Note that for all trials, the gauge was pre-filled with 50.

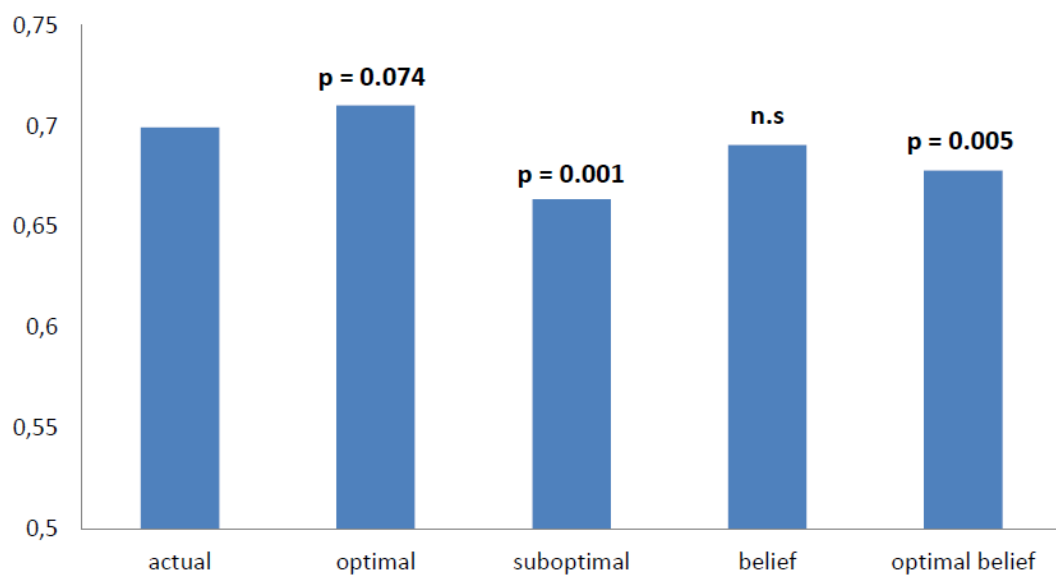


Fig. S3. Predicted Success. Comparison between mean actual success of groups (69.9%) and prediction of the different models: optimal (71.1%), suboptimal (66.4%), belief (69.0%) and optimal belief (67.8%). The p-values show that only the belief model is not statistically significant from the actual success.