Collective implementation of the Fermi method underlies the wisdom of deliberative crowds

Keywords: collective decision making, wisdom of crowds, natural language processing, word embeddings, Fermi estimation method

Extended Abstract

Understanding the conditions under which groups outperform individuals is a key problem in the social and psychological sciences. A central concept in this field is the “wisdom of crowds” principle, which suggests that the aggregation of independent estimates can often be more accurate than the best individual judgement (2). This phenomenon has been documented in various domains, including politics, health, and business, where it has been applied to forecast presidential elections (3), optimize cancer detection (4), improve leadership in organizations (5), and predict epidemic diseases (6). Previous studies have shown that small groups' deliberation and social influence can enhance crowd estimations (7). For example, it has been found that combining collective answers after group discussions results in more accurate estimations than averaging the initial independent responses (8). However, despite recent efforts in understanding the effect of social influence on the wisdom of crowds, the mechanisms by which deliberation increases the crowd’s accuracy remains unknown.

The present work fills that gap by examining the deliberative procedures that groups implement to increase their accuracy in decision-making. We hypothesized that groups use a technique called the "Fermi Method" (9) to decompose complex questions into simpler ones and make estimations based on limited or uncertain information and that, by doing so, they improve the precision of their estimations.

To test this idea, we performed a first study with N=588 participants, where we asked them to respond to general knowledge questions (e.g., "How many people live in South America?”). In the first stage, the participants answered eight questions individually. In the second stage, subjects were randomly assigned to 147 groups of four people each and were invited to a chat room where they had to discuss four of the questions (randomly selected from the initial eight questions) and arrive at a group consensus. In the final stage, they provided revised individual estimates for the eight questions (**Figure 1**).

To analyze the deliberation procedures and to identify whether the Fermi method was employed by the groups, we examined the chat logs using three different procedures with varying degrees of human supervision:

i) Supervised method: for this approach, 10 raters (who were naïve to the hypotheses of the study) were trained on what was the “Fermi method”. Then, they individually read each of the chat conversations and judged how much the Fermi method was used in a scale from 0 (not at all) to 10 (definitely used). They were also asked to rate how much the groups implemented a different strategy that consisted of exchanging their answers and combining their initial numerical estimates to obtain a final group answer (a procedure we call the “Numbers” strategy).

ii) Semi-supervised method: We analyzed the chat transcripts automatically and calculated the frequency of use of the words related to the question as a proxy for Fermi method use. As groups break down the question into smaller problems, they discuss related questions. For instance, to calculate the population of South America they may discuss about how many countries there are and what’s each of those countries’ population, increasing the frequency of words like “Brazil”, “Argentina”, “population”, “countries”, etc. (versus only mentioning numbers when they exchange their individual answers in a “Numbers” strategy). The related words were defined a priori by the authors for each of the questions in a “Whitelist.”

3. Unsupervised method: We analyzed the chat logs and represented the words mentioned in the discussions and in the questions in space vectors based on the word's meaning similarity (how semantically close they are from each other). We then calculated the distance (or similarity) between the chats and questions.

The results of Study 1 replicated previous results, showing that the average of the group's estimates is more accurate than the mean of the independent initial estimates (Cohen's d = .82, paired t-test, t(146)=9.9, $p=7\times10^{-18}$). Beyond this, we found evidence that the groups that implemented the Fermi method (as evaluated through any of the three methods) had lower error in their responses and were closer to the correct answer (**Figure 2**).

One main limitation of the results of Study 1 is that they are preliminary and observational in nature. To examine the replicability and validity of the results, we conducted a second study (N=224) in which we randomly assigned participants to two conditions. In one condition, participants were instructed to break down the problem in smaller steps and make the calculations required to reach to a reasonable estimation (i.e., they were explicitly instructed to apply the Fermi method). In the second condition, they were instructed to share their individual answers and combine those numbers to estimate the final answer to the question (i.e., they were asked to share numbers and aggregate those numbers to reach a collective estimate). Overall, we collected data from 56 groups of four people each (28 groups per condition). The results of this study showed that groups using the "Fermi" method reached to more accurate responses than groups that exchanged numbers. Moreover, we also observed that the groups using the Fermi procedure performed better than the groups that were not instructed to follow any particular strategy in Study 1. These findings provide causal evidence that using Fermi method for group discussions improves estimations.

Overall, this work shows that it is possible to improve the methods of problem estimation under uncertainty and improve the wisdom of crowds in collective decision making. Our findings suggest that the implementation of the Fermi method, either by instructing participants to use it or by identifying its implementation through automatic (word-embedding) or semi-supervised (whitelisting) methods can lead to more effective problem-solving in group settings.

References

1. Larrick, R. P., Mannes, A. E., Soll, J. B., & Krueger, J. I. (2011). The social psychology of the wisdom of crowds. Social psychology and decision making, 227-42.
2. Surowiecki, J. Te Wisdom of Crowds (Little, Brown, London, 2004)
3. Forsythe, R., Nelson, F., Neumann, G. R. & Wright, J. Anatomy of an experimental political stock market. Am. Econ. Rev. 82, 1142–1161 (1992)
4. Kurvers, R. H. et al. Boosting medical diagnostics by pooling independent judgments. Proc. Natl Acad. Sci. USA 113, 8777–8782 (2016).
5. Matzler, K., Strobl, A., & Bailom, F. (2016). Leadership and the wisdom of crowds: How to tap into the collective intelligence of an organization. Strategy & Leadership, 44(1), 30-35.
6. Li, E. Y., Tung, C. Y., & Chang, S. H. (2016). The wisdom of crowds in action: Forecasting epidemic diseases with a web-based prediction market system. International Journal of Medical Informatics, 92, 35-43.
7. Gürçay, B., Mellers, B. A. & Baron, J. Te power of social infuence on estimation accuracy. J. Behav. Decis. Mak. 28, 250–261 (2015).
8. Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nature Human Behaviour, 2(2), 126-132
9. Nityananda, R. (2014). Fermi and the art of estimation. Resonance, 19(1), 73-81.

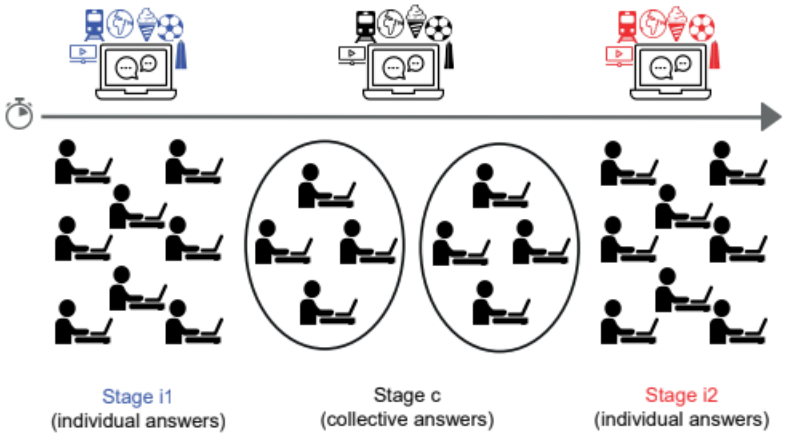

Figure 1. Procedure: in the first stage, subjects had to answer general knowledge questions. In the second stage, they were randomly assigned to groups of four to discuss the questions in a chat room and provide a group answer. In the final stage they provided their revised individual estimates for the questions.

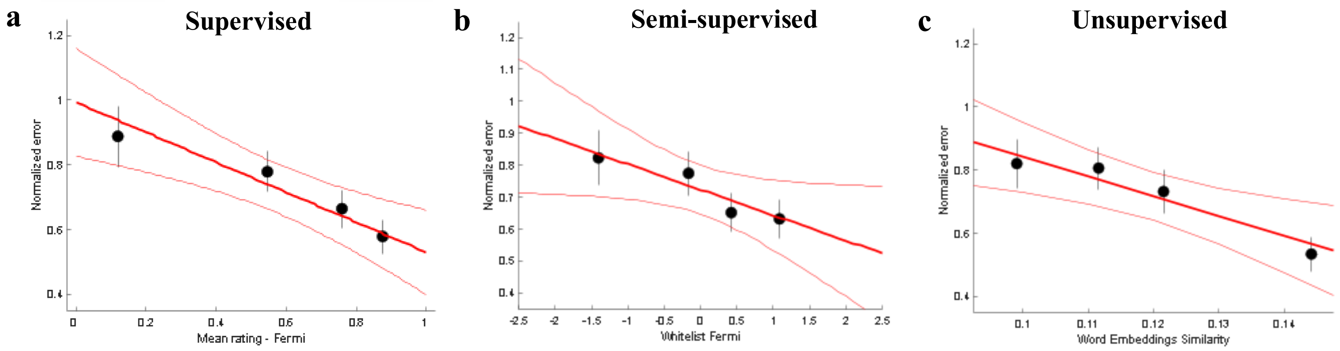

Figure 2. The normalized error (distance to the correct answer) decreases as Fermi method becomes more heavily used. In each panel, dots show the binned scatter plots using quartiles in the distribution of each variable in the x-axis. The red lines show the best-fitting linear model of the original data with their 95% confidence intervals. **a.** supervised method (β=-.14±.04, p=.0004), **b.** semi-supervised method (β=-.08±.04, p=.04), and **c.** word embeddings method (β=-.11±.04, p=.003).