The wisdom of crowds in one mind: Experimental evidence on repeatedly asking oneself instead of others

Heiko Rauhut
Chair of Sociology, in particular of Modeling and Simulation, ETH Zurich, Universitätstraße 41, 8092 Zurich, Switzerland, rauhut@gess.ethz.ch,

Jan Lorenz
Chair of Systems Design, ETH Zurich, Kreuzplatz 5, 8032 Zurich, Switzerland, jalorenz@ethz.ch, post@janlo.de

Under the right circumstances, groups can be remarkably intelligent and statistical aggregates of individuals' decisions can outperform individual's and expert's decisions. Examples of this wisdom of crowds effect range from markets, auctions, political polls, internet search engines to quiz shows (Galton, 1907; Lorge, Fox, Davitz, & Brenner, 1958; Surowiecki, 2004; Page, 2007; Mannes, 2009). Recently, Vul and Pashler (2008) and Herzog and Hertwig (2009) demonstrated a wisdom of crowds effect *within one mind* by an experiment, in which individuals could respond to the same question a second time. The underlying idea is that individual estimates are draws from an internal probability distribution such that their different estimates represent answers derived from different arguments or bodies of knowledge (Vul & Pashler, 2008; Stewart, 2009), relating to the intuitive idea of sleeping on difficult decisions.

Vul and Pashler (2008) demonstrated that delaying the second request for three weeks improved its value. Herzog and Hertwig (2009), by prompting individuals to consider knowledge that was previously overlooked, or deemed inconsistent (consider-the-opposite technique). While this evidence suggests that individuals are able to simulate a crowd in their brain, it remains open how beneficial this simulated crowd can become. We extend the previous work by eliciting five consecutive estimates from the same individuals. This extension allows to analyze how many own guesses are equivalent to asking someone else and to approximate the gained value of asking oneself ad infinitum by extrapolating the time trend to the limit. Further, we demonstrate that the formal treatment of Vul and Pashler (2008) does not generalize from two to multiple individual responses and suggest an improved measurement by computing an *average individual bias*.

This paper benefited from comments of Dirk Helbing, Michael Mäs, Ryan Murphy and Frank Schweitzer, and from research assistance by Hanna Thorn and Silvana Jud. We thank the ETH Zürich Competence Center 'Coping with Crises in Complex Socio-Economic Systems' (CCSS) through ETH Research Grant CH1-01-08-2 and the ETH Foundation for partial support. Address correspondence to rauhut@gess.ethz.ch or post@janlo.de

Method

We recruited 144 participants from ETH Zürich and conducted twelve sessions with twelve subjects each. All participants were asked to provide five consecutive answers without any information about the other subjects' estimates. Six different estimation tasks were probing their real-world knowledge, such as "What is the population density in Switzerland?" or "How many murders were registered in Switzerland in 2006?". Subjects received monetary payments taking into account the distance between estimate and true value ($0-10$ % (1.40 CHF), $11-20$ % (0.70 CHF), $21-40$ % (0.35 CHF), > 40 % (0 CHF)). This induced truthful revelation of judgements, resembling a scoring rule (Camerer, 1995). The order of questions was randomized across sessions. This procedure delivered the sample $\tilde{x}$, consisting of 1440 raw data points ($N_t = 5$ time steps, $N_i = 48$ subjects per question, $N_q = 6$ questions) with $\tilde{x}^q_{i,t}$ describing the t-th answer of subject i to question q.

Results

For calculating the wisdom of crowds, we aggregate the individuals' estimates to the geometric mean. This takes into account that the distribution is skew and non-Gaussian and delivers aggregate results closer to the truth than the standard arithmetic mean. Therefore, we normalized and transformed the raw data $x^q_{i,t} = \log \frac{\tilde{x}^q_{i,t}}{\text{truth}^q}$ so that the true values correspond to zero and the arithmetic mean of x delivers the logarithm of the geometric mean of $\tilde{x}$. Our comparison of asking repeatedly oneself with asking somebody else is represented by comparing the sample x with the sample z, generated by randomly sampling initial estimates $z^q_{j,t} = x^q_{\text{rand},1}$. We bootstrapped $z^q_{j,t}$ for $j = 1,\dots,144000$ repetitions. Thus, $z_{j,1},\dots,z_{j,5}$ represents estimates from five different individuals, while $x_{i,1},\dots,x_{i,5}$ represents five estimates from the same individual.

The wisdom of the crowd refers to a comparison between the belief of the crowd and the truth. We define the *belief of the crowd* of T estimates of question q and sequence i as $\text{BoC}^q_{i,T}(x) = \frac{1}{T}\sum_{t=1}^T x^q_{i,t}$. The wisdom of the crowd can be expressed by the difference between belief and truth, represented by the mean squared errors $\text{MSE}_T(x) = \frac{1}{N_q N_i}\sum_{q,i}\left(\text{BoC}^q_{i,T}(x) - \text{truth}^q\right)^2$. This applies to both samples x and z, allowing direct comparisons between asking

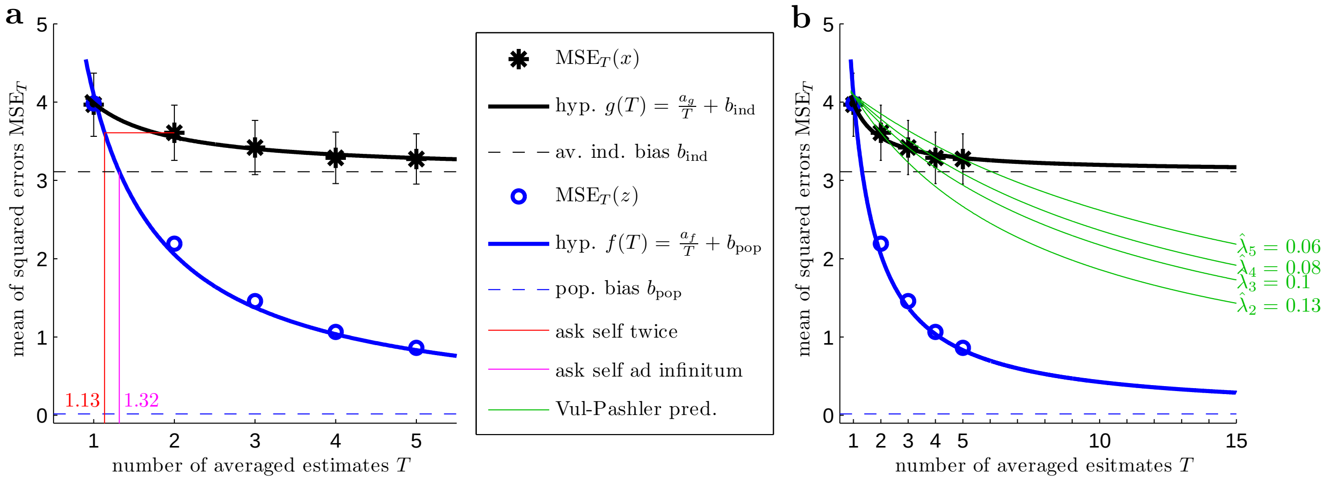

Figure 1. The wisdom of crowds for asking repeatedly oneself compared to asking others (144 subjects, 288 responses). (a) The x-axis denotes T answers from the same respectively different individuals. The y-axis represents the wisdom of crowds, measured by the mean of the squared errors MSE_T between estimates and truth. The blue circles denote the wisdom of asking different individuals ($\mathrm{MSE}_T(x)$) and the black stars of asking repeatedly oneself ($\mathrm{MSE}_T(z)$). The error bars at the stars correspond with the 5 % bootstrapped confidence intervals; the confidence intervals for different individuals lie within the blue circles. The dashed blue line represents the error for asking all different individuals and the dashed black line for asking oneself ad infinitum. Asking oneself twice corresponds with asking 1.13 other individuals (red line) and asking oneself ad infinitum corresponds with asking 1.32 other individuals (magenta line). (b) Demonstration of the alternative method of Vul and Pashler, extended to 15 responses for illustrative purposes. Vul-Pashler hyperbolas are shown as green lines with λ's estimated for $\mathrm{MSE}_2(x)$, $\mathrm{MSE}_3(x)$, $\mathrm{MSE}_4(x)$ and $\mathrm{MSE}_5(x)$. Their parameter λ for fitting the hyperbola does not reproduce the experimental results and underestimates the error of asking oneself ad infinitum.

oneself and others.[1]

The gained wisdom of asking different individuals requires to fit the $\mathrm{MSE}_T(x)$ over asking $T = 1,2,3,4,5$ other individuals as a hyperbola with reference to the limiting *population bias* $b_{\mathrm{pop}} = \left(\frac{1}{N_q N_i}\sum_{i,q}(x^q_{i,1} - \mathrm{truth}^q)\right)^2$. Because b_{pop} expresses the error of the averaged estimates, it allows a compensation of over- and underestimates for all estimates, while $\mathrm{MSE}_T(z)$ limits the benefit of diversity to T estimates. We fit the hyperbola $f(T) = \frac{a_f}{T} + b_{\mathrm{pop}}$ by estimating the parameter a_f ($R^2 = 0.99$).[2]

Analogously to Vul and Pashler (2008), we compare the error of asking oneself T times with asking T different people by projecting $\mathrm{MSE}_T(x)$ horizontally to f and down to the x-axis. The red line in Figure 1a demonstrates that asking oneself two times corresponds with asking 1.13 different persons. This finding resembles the value of 1.11 of Vul and Pashler (2008) in their "non-delayed condition".

In the following, we extend the analysis of Vul and Pashler (2008) from asking oneself twice to multiple times. Given the theory that individual estimates are draws from a probability distribution, the average of several estimates converges asymptotically to the average individual bias $\mathrm{MSE}_T(x)$ as a hyperbola $g(T) = \frac{a_g}{T} + b_{\mathrm{ind}}$. Note that both, a_g and b_{ind} are fitted ($R^2 = 0.98$) because our number of estimates from the same individual is restricted to five. The average individual bias represents the remaining average error when individuals ask themselves ad infinitum. By projecting the average individual bias horizontally to the hyperbola f and down to the x-axis (magenta line in Figure 1a), we find that asking oneself ad infinitum can only yield as much reduction of mean squared error as asking 1.32 different persons.

Our method of estimating an individual bias is more general than the proposal of Vul and Pashler (2008), because it is not restricted to two time steps. They estimate λ, which is "the proportion of an additional guess from another person that an additional guess from the same person is worth". Asking oneself T times should thus correspond to asking different people $1 + \lambda(T-1)$ times, which implies that the $\mathrm{MSE}_T(x)$ follows the hyperbola $h(T) = \frac{a_f}{1+\lambda(T-1)} + b_{\mathrm{pop}}$. Figure 1b demonstrates that there is no globally valid lambda. The method overestimates the benefit of repeatedly asking oneself and does not reproduce the experimental data. The problem is that their method neglects the existence of an individual bias different from the population bias, which bounds the benefit of asking oneself repeatedly.

Discussion

Our data demonstrates that asking oneself repeatedly can improve the estimation of vaguely known facts; even without information feedback from others. The effect is, however, not as strong as previous analyses suggested. While asking oneself two times is as good as asking 1.13 other individuals, asking oneself ad infinitum does not improve to more than

[1] Note that $\mathrm{truth}^q = 0$ in x and z.

[2] The hyperbola is used due to the central limit theorem.

1.32 representative others. Refined analyses showed that these values vary across different types of questions. Easier questions revealed that asking oneself several times can measure up to asking 2.5 different persons, while the effect almost vanishes for more difficult questions. Further, if questions are emotionally loaded, we even observed worsened results.

References

Camerer, C. (1995). Individual decision making. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics.* Princeton, NJ: Princeton University Press.

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

Herzog, S., & Hertwig, R. (2009). The Wisdom of Many in One Mind: Improving Individual Judgments With Dialectical Bootstrapping. *Psychological Science*, *20*(2), 231–237.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group-performance and individual-performance, 1920-1957. *Psychological Bulletin*, *55*(6), 337–372.

Mannes, A. E. (2009). Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science*, 1-13.

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools and societies*. Princeton University Press.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology*, *62*(6), 1041-1062.

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.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*.