CellPress

# Harnessing the wisdom of the inner crowd

**Stefan M. Herzog and Ralph Hertwig**

Max Planck Institute for Human Development, Center for Adaptive Rationality, Lentzeallee 94, 14195 Berlin, Germany

**Ever since Galton's classic demonstration of the wisdom of crowds in estimating the weight of a slaughtered ox, scholars of the mind and the public alike have been fascinated by the counterintuitive accuracy achieved by simply averaging a number of people's estimates. Surprisingly, individuals can, to some extent, harness the wisdom of crowds within the confines of their own mind by averaging self-generated, nonredundant estimates.**

## The powerful principle of judgment aggregation

Sir Francis Galton, the Victorian polymath, was an outspoken elitist and skeptical of the value of democratic judgment. In 1906, he attended a livestock fair at which visitors guessed the weight of a butchered and dressed ox. Although the 787 estimates varied widely, the median estimate was only 9 pounds (0.8%) off the mark. The remarkable precision of the *vox populi* surprised Galton: 'This result is (. . .) more creditable to the trust-worthiness of a democratic judgment than might have been expected' [1]. Since Galton's demonstration of the 'wisdom of crowds' [2], research in psychology, economics, political science, biology, statistics, and computer science has time and again demonstrated how aggregating diverse judgments frequently – and sometimes dramatically – increases accuracy because nonredundant errors cancel each other out. Surprisingly, a lone individual can also enlist the wisdom of crowds by averaging self-generated, nonredundant estimates [3–12]. We review evidence for this 'wisdom of the inner crowd', and consider how it can be produced, how its accuracy can be improved, and whether people use it to their advantage.

## How wise is the inner crowd?

Stroop first demonstrated the wisdom of the inner crowd in the 1930s [10]. His participants ranked identical-looking weights according to their heaviness, with some participants returning up to 50 times to rank the same set of weights anew. When Stroop averaged across ever more rankings originating from either different people or the same person, he found that the correlation with the true ranking of the weights increased to the same extent for both real and inner crowds. Because the weights looked identical, participants were not able to remember their rankings from previous sessions. The errors of repeated rankings by the same participants were thus as independent as those by different participants. Stroop's results illustrate that, under this ideal condition, inner crowds can be as powerful as real crowds.

Having lain dormant for some time, the wisdom-of-the-inner-crowd effect has recently been rediscovered. The benefits of averaging multiple estimates provided by the same person have been demonstrated for various general knowledge quantities (proportions [3,5,6,8,11], historical dates [4,8], and correlations [12]). However, evidence of averaging benefits is somewhat mixed for both general knowledge questions without clear-cut boundaries for reasonable answers [8,9] and estimates of social consensus (a correlational accuracy measure showed no benefits, but a deviation-based accuracy measure did) [7], and the only study investigating the quality of confidence judgments found no benefits of harnessing the inner crowd [13]. In summary, most – but not all – studies have found evidence for the wisdom of the inner crowd. Some of the differences in results may stem from methodological differences (e.g., in how averaging was implemented or how accuracy was measured); others may be due to the knowledge domains and types of judgments investigated. Future research is needed to understand the boundary conditions of the wisdom of the inner crowd (see Box 1 for one approach).

With the exception of Stroop's study [10], research has found real crowds to produce larger averaging gains than equally sized inner crowds [4,8,9,11–13]. The reason is that their errors are less redundant than are those observed in inner crowds [4]. In real crowds, returns therefore diminish at a slower rate when more estimates are added [9,12], because error redundancy imposes an upper ceiling on how much and how quickly accuracy can be increased by error cancellation [9]. Adding another person is thus likely to generate more benefit than is asking oneself once more [12].

## How to foster the wisdom of the inner crowd?

Given that reducing error redundancy should increase averaging gains within a person, an obvious question is this: What reduces error redundancy within a person? The key insight is that error redundancy reflects, among other things, the amount of redundant information contained in multiple estimates – either by the same person or by different people [6,13]. Consistent with this, averaging gains are larger for people with lower (vs higher) working memory spans [6]. These estimates are probably based on a smaller sample of information retrieved from memory; smaller samples randomly drawn from the same underlying distribution are necessarily less redundant than larger samples. Error redundancy also decreases when previous estimates are not permitted to hold strong sway over subsequent ones. Specifically, techniques that reduce the 'control' of previous estimates can reduce error redundancy

*Corresponding author:* Herzog, S.M. (herzog@mpib-berlin.mpg.de).
*Keywords:* estimation; judgments under uncertainty; inner crowd; crowd within; dialectical bootstrapping; judgment aggregation; wisdom of crowds.

1364-6613/
 http://dx.doi.org/10.1016/j.tics.2014.06.009

**Box 1. When to enlist the inner crowd**

The probability, accuracy, redundancy (PAR) model [15] was developed to study advice taking (i.e., with one judge and one advisor). However, it also offers a framework to investigate the ecological conditions under which averaging one's self-generated estimates outperforms the strategy of betting on the more accurate set of self-generated estimates [3,5]. The model's parameters are:

- The probability **p** of identifying the better set of estimates, that is, a person's skill to correctly predict whether, overall, the first or second estimates are more accurate.
- The accuracy ratio **A**, which is defined as the ratio of the errors of both sets of estimates (larger error divided by smaller).
- The redundancy of the errors operationalized as the 'bracketing rate' **Br**, that is, the proportion of questions for which the answers have errors of different signs and thus errors cancel each other out (with higher Brs indicating lower error redundancy).

Figure I shows a PAR model analysis of the wisdom of the inner crowd [5]. The PAR model illustrates that when Br is high enough (i.e., error redundancy is low enough), always averaging is more accurate than is choosing the better set of estimates, even with perfect skill (p = 1). Data are from a study [5] that compared a dialectical-bootstrapping condition (using the consider-the-opposite technique) with a reliability condition (in which participants re-estimated a quantity without special instructions). In both conditions, the first and second sets of estimates were similarly accurate (i.e., both A values were close to 1) and had small to moderately large bracketing rates. Most importantly, the dialectical condition was located slightly above the iso-accuracy curve for p = 1; that is, even someone who always detects the better set of estimates would not outperform the wisdom of the inner crowd. The only way to do so would be to abandon averaging on a trial-by-trial basis under the appropriate circumstances; however, people do not seem able to accurately identify those circumstances [3,5].

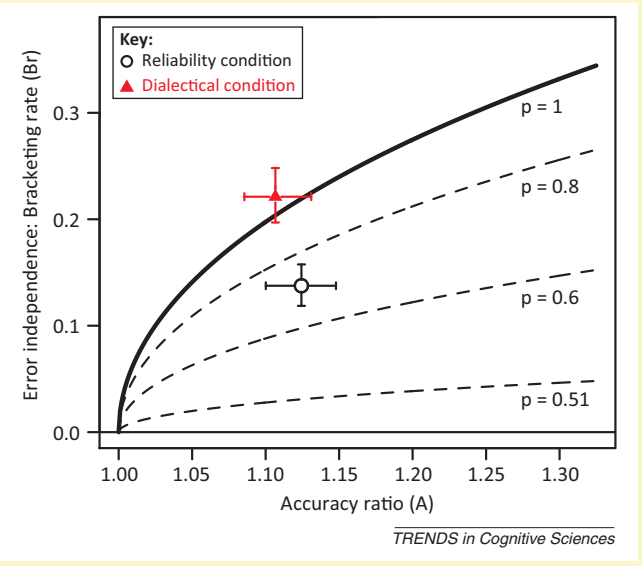


**Figure I.** The cognitive–environmental niche for the wisdom of the inner crowd. The lines for different values of p represent iso-accuracy curves, that is, combinations of accuracy ratios (A) and bracketing rates (Br) for which averaging the estimates versus choosing the better set of estimates (with skill p) return the same accuracy. For combinations above an iso-accuracy curve, averaging is better than choosing with skill p; for combinations below, choosing is better. Data are from [5]; error bars show 95% highest density intervals (HDIs). Adapted from [5].

and boost averaging gains [4,11]. At least two classes of such techniques are available.

One approach capitalizes on the power of forgetting. For example, increasing the time delay between subsequent estimates – and thus freeing latter estimates from the traction forces of former estimates – can increase averaging gains [11]. A second approach takes advantage of the mind's ability to construct alternative realities. People can be actively prompted to base their second estimate on different assumptions, pieces of information, inference methods, and elicitation procedures [4,12]. One way to achieve this process of 'dialectical bootstrapping' [4,5,14] is to ask people to 'consider the opposite' when making a new estimate. This prompt has been shown to decrease error redundancy and increase averaging gains [4,5] (but see also [14]).

## Do people harness the wisdom of their inner crowd?

People underestimate the benefit to be gained from averaging their own opinions with those of advisors [2,15]. But do they appreciate the potential gains of taking their own 'advice'? Indeed, people asked to give a final estimate in light of their self-generated previous ones tend to combine them (i.e., to give a final estimate located somewhere between the first two) [5,8], although few people strictly average their estimates by giving equal weight to both of them [5,8]. Moreover, when explicitly asked to decide between choosing their first estimate, choosing their second estimate, or averaging both estimates, decision makers most frequently opt for averaging [3].

People are more likely to combine their estimates if they actively challenge the premises of their first answer in the process of generating a second one [5]. Prompting people to adopt more than one perspective on the same problem – to 'consider the opposite' – may make them aware of conflicting, yet legitimate, assumptions and reasons. Combining the resultant estimates is an elegant tool for trading off these conflicting realities. Another factor conducive to combination is the magnitude of disagreement between two estimates. The larger the numerical difference between the two estimates, the more inclined people are to combine them [5]. One interpretation is that they are especially motivated to hedge against the risk of selecting the wrong estimate when disagreements are large.

Although people generally tend to combine their self-generated estimates, they also often choose among them [3,5,8]. Do they abandon combining under the right circumstances? To date, the research indicates that this is not the case. People seem unable to outperform the inner crowd and would be better off always strictly averaging their estimates [3,5,8]. Although people tend to select the more accurate of their estimates when choosing between them [3], this accuracy in spotting the better estimate does not suffice to outperform the averaging of estimates [3,5]. Averaging is not superior in every environment [3,5,15] (see Box 1 for an explanation), but in the environments investigated so far, the accuracy of the inner crowd was unsurpassed.

## Concluding remarks: how to exploit the inner crowd

Individuals can boost the accuracy of their estimates by consulting an outer crowd or enlisting an inner crowd (Figure 1). With regard to the inner crowd, several challenging questions remain to be answered. First, the boundary conditions of this mental device need to be delineated

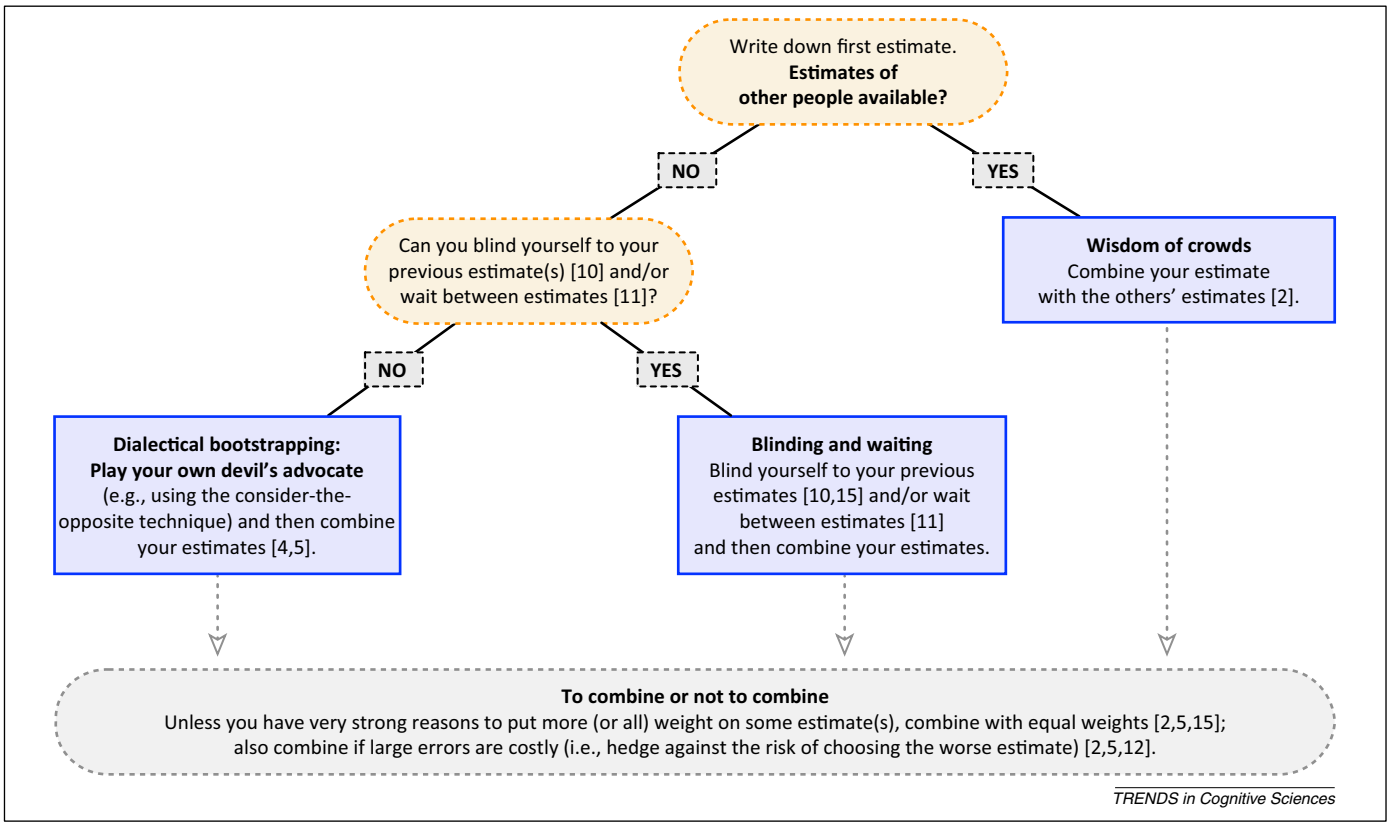


**Figure 1**. Decision tree for deciding when and how to use the inner crowd.

(e.g., can it also be generalized to judgments that are not quantitative?). Second, research can help people to exploit their inner crowd by identifying further tools that can reduce error redundancy and increase the propensity to harness the inner crowd under the appropriate circumstances (Box 1). Three tools have already been found – time delay [11], consider-the-opposite technique [4,5], and different methods of assessment [12] – but other easy-to-use techniques are likely to exist. A third open question concerns the conditions under which people spontaneously generate more than one opinion. The fourth challenge is to identify people who could benefit more than others from enlisting the inner crowd. For instance, do elderly people with declining cognitive resources benefit more (akin to people with lower working memory span [6])?

The surprising accuracy of the *vox populi* challenged Galton's elitist weltanschauung. Likewise, the wisdom of the inner crowd rehabilitates the existence of conflicting thoughts and opinions within a single mind. Rather than striving to identify one 'superior' response, decision makers can benefit from actively consulting different opinions within their mind and harnessing their wisdom by aggregation.

## Acknowledgments

We thank Susannah Goss for editing the manuscript and the Swiss National Science Foundation for a grant to both authors (100014_129572/1).

## References

1 Galton, F. (1907) Vox populi. *Nature* 75, 450–451
2 Larrick, R.P. *et al.* (2012) The social psychology of the wisdom of crowds. In *Frontiers in Social Psychology: Social Judgment and Decision Making* (Krueger, J.I., ed.), pp. 227–242, Psychology Press
3 Fraundorf, S.H. and Benjamin, A.S. (2014) Knowing the crowd within: metacognitive limits on combining multiple judgments. *J. Mem. Lang.* 71, 17–38
4 Herzog, S.M. and Hertwig, R. (2009) The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychol. Sci.* 20, 231–237
5 Herzog, S.M. and Hertwig, R. (2014) Think twice and then: combining or choosing in dialectical bootstrapping? *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 218–232
6 Hourihan, K.L. and Benjamin, A.S. (2010) Smaller is better (when sampling from the crowd within): low memory-span individuals benefit more from multiple opportunities for estimation. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1068–1074
7 Krueger, J.I. and Chen, L.J. (2014) The first cut is the deepest: effects of social projection and dialectical bootstrapping on judgmental accuracy. *Soc. Cogn.* 32, 315–335
8 Müller-Trede, J. (2011) Repeated judgment sampling: boundaries. *Judgm. Decis. Mak.* 6, 283–294
9 Rauhut, H. and 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
10 Stroop, J. (1932) Is the judgment of the group better than that of the average member of the group? *J. Exp. Psychol.* 15, 550–562
11 Vul, E. and Pashler, H. (2008) Measuring the crowd within: probabilistic representations within individuals. *Psychol. Sci.* 19, 645–647
12 Winkler, R.L. and Clemen, R.T. (2004) Multiple experts vs. multiple methods: combining correlation assessments. *Decis. Anal.* 1, 167–176
13 Ariely, D. *et al.* (2000) The effects of averaging subjective probability estimates between and within judges. *J. Exp. Psychol. Appl.* 6, 130–147
14 Herzog, S.M. and Hertwig, R. (2013) The crowd-within and the benefits of dialectical bootstrapping: a reply to White and Antonakis (2013). *Psychol. Sci.* 24, 117–119
15 Soll, J.B. and Larrick, R.P. (2009) Strategies for revising judgment: how (and how well) people use others' opinions. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 780–805