Extracting the Wisdom from the Crowd: A Comparison of Approaches to Aggregating Collective Intelligence

Thomas Görzen[1], Florian Laux[1]

[1]Paderborn University, Business Administration and Economics, Paderborn, Germany
{Thomas.Goerzen,Florian.Laux}@wiwi.uni-paderborn.de

Abstract.
To benefit from crowdsourcing, companies are increasingly required to employ mechanisms for aggregating the multiple opinions generated in this process. Previous research, however, has raised concerns with the currently most popular method used for this purpose: majority voting. We conduct an experiment to compare different aggregation methods and measure their performance. Our results confirm these concerns and identify other, confidence-based aggregation approaches that provide significantly better results in identifying the right answer. Moreover, by differentiating between different levels of question difficulty, we find that the average confidence approach provides the highest percentage of correctly identified answers across different categories of questions. Our findings both extend the existing literature on aggregation approaches used for collective intelligence, and offer practical insights. Since we use a crowd on a commercial crowdsourcing platform, our results offer valuable insights for companies using or planning to use a crowd for collective intelligence.

Keywords: Crowdsourcing, collective intelligence, wisdom of the crowd, aggregation approaches.

1 Introduction

Crowdsourcing is a large and growing phenomenon [1]. It is defined as "[…] *the act of taking a task once performed by an employee and outsourcing it to a large, undefined group of people external to the company in the form of an open call*" [2]. One application of crowdsourcing is so-called crowd voting, in which a large number of people are recruited on online platforms to give their opinion on various topics [3]. Applications of crowd voting include political and economic forecasting, public policy, or evaluating nuclear safety [4]. Firms also increasingly employ mechanisms for aggregating multiple opinions, especially when navigating markets that are difficult to predict (e.g., [5], [6]).

Multikonferenz Wirtschaftsinformatik 2018,
March 06-09, 2018, Lüneburg, Germany

The concept of crowd voting is generally based on the wisdom of the crowd effect, which posits that the aggregated opinion of the crowd is superior to any individual opinion, even those of experts [7]. However, one major challenge when drawing on the wisdom of the crowd is how to aggregate the different, heterogeneous opinions for the best and most meaningful overall results. Currently, the most popular way to aggregate a multitude of opinions is a democratic voting procedure, in which all opinions are simply aggregated without using any weighting or filtering of judgements [8]. However, this method has serious limitations. The literature finds that, for example, individuals' judgements are frequently too extreme, that they are overconfident in their reported rating ability, or that their judgments are biased by the anchor effect [9], [10]. Another reason why crowds' wisdom might fail is because the aggregate estimate was largely distorted by a systematic group bias or by a large number of uninformed judges [11]. This leads us to the following research question:

Which approach for extracting wisdom from the crowd yields the best and most reliable results in the context of crowd voting?

Previous studies suggest different approaches for aggregating multiple judgements, including a new and promising alternative approach, the so-called *surprisingly popular approach*. As the name suggests, this approach is based on selecting the answer that is unexpectedly more popular than people predict it to be [4]. While this approach has haled to good and reliable results in one prior study [4] which used selected offline crowds (students and dermatologists), it has not yet been tested with a real-life anonymous online crowd. Hence, our study aims to investigate whether the *surprisingly popular approach* also leads to good and reliable results when used with a more diverse, anonymous online crowd on a commercial online platform. We then compare this new approach with other aggregation mechanisms used in this context to investigate which approach provides the best and most reliable results.

2 Background

2.1 Wisdom of the Crowd

The wisdom of the crowd effect has a long history and has attracted the attention of scholars for a considerable time (e.g. [12], as an early example) including current researchers (e.g. [13]). The general idea of this effect is simple and can be described as follows: when predicting an unknown outcome (for example the weight of an object), the central tendency of different, individual estimates represents the true value of the unknown outcome more closely than any one individual estimation [14]. In comparison to a single individual estimation, this approach offers several advantages because it i) maximizes the amount of information available for the decision, estimation, or prediction task; ii) reduces the potential impact of extreme or aberrant sources that rely on faulty, unreliable, and inaccurate information; iii) increases the credibility and validity of the aggregation process by making it more inclusive and ecologically more representative [15]. Previous literature [16] suggests that there are conditions that are

necessary for a crowd to be wise: i) the diversity of the crowd, ii) a particular kind of decentralization and iii) the independence of judgements. In the online contexts, the third condition especially is often violated since, if information is provided, people can easily observe the decision made by others, leading to decision making influenced by a previous decision maker [17]. Literature demonstrated that even a mild social influence can undermine the wisdom of the crowd effect for simple estimation tasks [18]. However, if the conditions described above are met, previous literature suggests that the average of the judges - the wisdom of the crowd effect - beats the average judge [8]. Estimations derived from a large heterogeneous group has even been found to outperform samples of homogeneous experts [7], [18].

2.2 Aggregation Mechanisms

The reason why groups outperform individuals is based on the statistical principle that aggregation of imperfect estimates reduces error, resulting in better and more reliable results [19], [20]. As mentioned in the introduction, due to the popularity of crowdsourcing, the methods used to aggregate multiple opinions for collective decision making has come under scientific scrutiny (e.g. [14]), in particular the popular democratic method which merely calculates the average of judgments. Its limitations include cognitive decision biases and lack of relevant knowledge on which individuals in the crowd base their decision-making [11], [21]. To overcome this limitation, different additional approaches have been developed (Table 1), including: Majority voting (MV), confidence weighted (CW), confidence only (CO), average confidence (AC), and surprisingly popular (SP). In the following, we briefly explain each of these approaches.

Table 1. Different aggregation mechanisms

	MV	CW	CO	AC	SP
Input	Multiple judgements	Multiple judgements, confidence		Confidence of the judgements	Multiple judgements, predicted popularity
Key numbers	Multiple judgements	Weighted number judgements	Number of judgements with confidence = 100%	Average confidence	Actual and predicted popularity
Decision criteria	Majority of judgements	Majority of weighted judgements	Majority of 100% confident judgements	Highest average confidence	Actual popularity > predicted popularity
Literature	[22]			[23]	[4]

Majority voting (MV) is the simplest approach to aggregate multiple judgements. The only input needed are the judgements of the crowd. The decision for one possible answer is based on the majority of votes for this answer.

The *confidence weighted* (CW) approach, in contrast, requires the confidence of each judgement as an additional input. By including the confidence of an answer it is possible for participants to signal their confidence level when answering the question.
Similarly to the confidence weighted approach, the *confidence only* (CO) approach also includes the confidence of each judgement. However, the confidence only approach solely includes judgements by participants with very high levels of confidence (confidence = 100%) in their ability to identify the correct answer. In contrast to the other approaches described above, the *average confidence* (AC) approach is solely based on the confidence of judgements to identify the correct answer. The decision criteria is based on the highest average confidence for a possible answer. In other words, this approach identifies the answer representing the highest average confidence as the correct answer.
The *surprisingly popular* (SP) approach is a relatively new approach, developed by Prelec at al. [4]. In contrast to other methods, this also takes into account the predicted popularity of an answer. Apart from asking each individual to judge which answer is the correct one, participants are then asked to predict what percent of the crowd they think will give a certain answer, for example, whether the statement is true. The decision for identifying the best answer is made by comparing the actual popularity of an answer given by the crowd with the predicted popularity. The algorithm identifies an answer as correct when its actual popularity is higher than the predicted popularity, hence the method's name, "surprisingly popular". To illustrate this approach, we use the following example: if we ask the crowd whether a saxophone is a brass instrument, we assume that the majority of judgements would answer "yes" since a saxophone is made of sheet metal. Hence, people who answer "yes" would also predict that most of the other individuals would give the same answer. However, a person with more knowledge in this topic area would know that the sound of a saxophone is generated by a wooden reed, which is why it is classified as a woodwind instrument. Therefore, this person would answer "no" and in addition, would also predict a high percentage of people who would answer with "yes" since she assumes that this specified knowledge is not widely shared.
Previous literature already investigated different approaches for collective decision making. Most related to our study is that by Prelec et al. [4] who compared the results of majority voting, surprisingly popular, confidence weighted and confidence only approaches. Their findings show that the surprisingly popular approach leads to the best and most reliable results [4]. However, participants used in this study have mainly been offline crowds (students) or homogeneous crowds (dermatologists), labeling specific images displaying benign and malignant lesions [4]. Therefore our study aims to compare the performance of five different approaches for aggregating multiple judgements by using a heterogeneous crowd on a commercial online crowdsourcing platform.

3 Experimental Design

To investigate which approach leads to the best and most reliable results, we designed an experiment on the online crowdsourcing platform *Crowdflower*. In line with

previous literature [4], we designed a task consisting of 35 factual questions of general knowledge, including the categories geography, music, literature, sports, politics and history. The questions were a subset of true/false quizzes from the quiz site *Sporcle*[1]. Since we used a commercial online crowdsourcing platform, the jobs offered here are usually business related tasks. To avoid that our tasks appear unnatural as well as to avoid experimenter demand effects [24], we told the subjects in our instructions that we work for an institute dealing with the general education of the population in Germany.

In a first step, we conducted a pilot experiment with the aim to differentiate the levels of question difficulty. We conducted this pilot experiment for two reasons: First, by differentiating between levels of difficulty we could ensure that each participant had to answer questions on a comparable level of difficulty. By so doing, we aimed to avoid subjects getting frustrated when facing too many questions they would find difficult. Second, the differentiation of difficulty enables us to further investigate how different approaches to aggregate the judgements perform for different levels of question difficulty.

We divided the questions into two blocks of 20 questions each, including five questions that were assigned to both blocks of ideas since we needed to have two equally sized blocks of 20 ideas each. We randomly assigned each participant to one of the two blocks of questions. Each contributor had to answer whether the proposed statement was right and if they answered that the statement was wrong, an additional answer field popped up, asking them to submit the right answer. Participants were not able to assign themselves to the task several times. Each participant received a monetary reward of 0.15€. In sum, 64 different subjects participated in this pilot. Based on the percentage of correct responses to the questions, we divided the questions into three different categories of difficulty (Table 2).

Table 2. Categories of difficulty

Categories	Percentage of correct answers	Number of questions
Easy	> 75%	13
Medium	75% - 50%	11
Difficult	< 50%	13

Based on the results of our pilot experiment, we designed the main experiment of our study. We formulated our instructions in line with our pilot experiment. Further, we adopted the type of questions to test the performance of the *surprisingly popular* algorithm by formulating each question as a statement where each subject had to decide whether the statement was right or wrong (in line with [4]). For example, one statement was: "A saxophone is a brass instrument" followed by the question, whether this statement is right or wrong. In addition, we also asked each subject how confident they felt in answering this question. Subjects could respond in integers varying from 50%, which is equivalent to a coin toss, meaning that they have been totally uncertain, to

[1] http://www.sporcle.com

100% which would indicate that they have been absolutely certain about their answer. We then asked each subject to think about other people's answer to this question and to predict the percentage of people who would give a certain answer, for example that the statement is true, varying from 1%, indicating that almost no other person said the statement is true, to 100%, indicating that all other persons would rate this statement as true. We informed the subjects that about 100 other subjects would answer the same questions to simplify the prediction of what percent of people would guess the answer as true (Figure 1). This design enables us to investigate different aggregation mechanisms since we monitor the majority voting of the crowd, the confidence levels of each question for the crowd, and the opportunity to apply the *surprisingly popular* algorithm by asking subjects about other people's belief in respect of the answer. Hence, we are able to analyze the results of five different aggregation mechanisms.

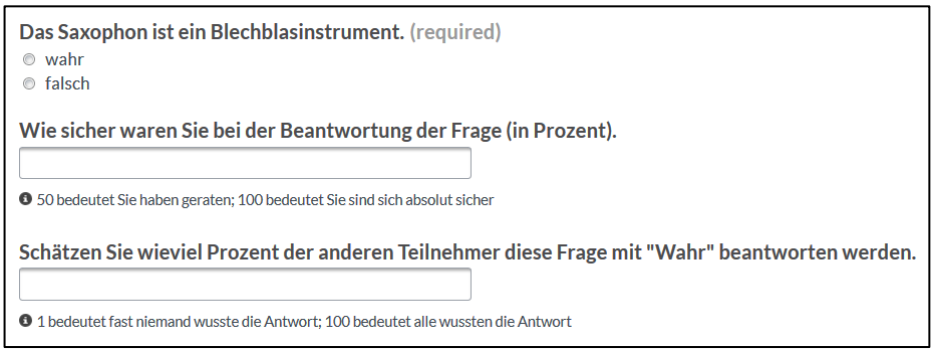

Figure 1. Example of a question

Since it would be unreasonable to ask each subject to answer all 35 questions, we split the questions into two blocks of 20 questions each based on our categories of difficulty. Because we could not divide 35 questions into two blocks of 20 ideas each, we equally filled the missing places with easy questions into both blocks to obtain equally sized question blocks. We added eight easy questions to each block. Since both categories of "medium" and "difficult" questions consist of odd numbers, we assigned six medium and six difficult questions into block one and five questions from the category "medium" as well as seven questions from the category "difficult" into question block two. We further developed a linear optimization model to compare the difficulties of questions in each block based on the exact percentage of correct answers given in the pilot experiment. The optimal solution found by the model excluded two questions from the category "easy". This seems acceptable, because all aggregation mechanisms should find the correct answer for these questions anyways. Finally, we controlled whether difficulties of questions within the first and the second half of each block of questions was comparable, in order to avoid, for example, assigning several difficult questions in succession.

In sum, 206 subjects (74.4% male) on average 37.9 years old, participated in our experiment, with 100 answering the questions in block 1, and 106 subjects answering the questions in block 2. Each participant received a monetary reward of 0.20€ for answering the block of questions. Participants also rated the clarity of instructions and

the payment on a five point scale, with five indicating the best possible value. As the clarity of instructions was rated as 4.3 and payment rated as 3.8, we concluded that participants were generally satisfied with the task design.

4 Results

4.1 Overall Performance

In a first step, we compare the overall performance of each approach without differentiating between difficulties of questions. Performance in this context is measured by the percentage of correctly identified answers when using different approaches to aggregate the judgements. Figure 2 shows that the *majority voting* approach provides the lowest percentage of correct answer (62%), while the *average confidence* approach provides the highest percentage of correct answers (97%).

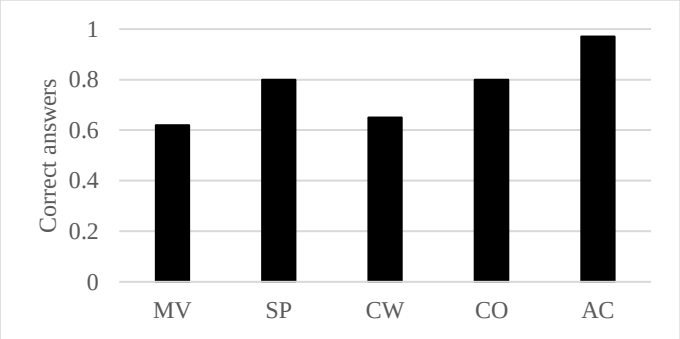

Figure 2. Overall performances

In a next step, we compared the uplift, i.e. the increase of performance between the different approaches, to aggregate the answers. Since the *majority voting* approach is the most commonly used approach in practice [8], and the overall worst performing approach, we calculated the differences between the performances of the *majority voting* and the other four approaches.
Following previous literature [4], we used two-sided matches-pair tests [25], [26] to compare the number of correct answers, hence the increase in performance of the different approaches compared to the *majority voting* approach.

Table 3. Increasing performance compared to majority voting approach

	Surprisingly Popular (SP)	Confidence Weighted (CW)	Confidence Only (CO)	Average Confidence (AC)
Perf. Uplift	+ 17.1%	+ 2.8%	+17.1%	+ 34.2%
P-Value	0.115	0.804	0.115	0.000

The *average confidence* approach generated the highest and the only statistically significant increase in overall performance compared to the *majority voting* approach (Table 3). However, taking into account the relatively small number of questions (n=40), the p-values for both the *surprisingly popular* as well as the *confidence only* approach come close to the allowable upper boundary of $p = 0.1$, indicating a significant increase compared to the *majority voting* approach.

Following previous literature [4], we further calculated classification accuracy by using categorical correlation coefficients such as Cohen's kappa [27] as well as Matthews correlation [28] which allows frequencies of different correct answers to be imbalanced, resulting in high percentage agreement driven by chance. In line with the results mentioned above, the *majority voting* approach shows the lowest percentage of agreement while the *average confidence* approach shows the highest agreement with the correct answers (Table 4). All levels of agreement for the different approaches are highly statistically significant, indicating that the agreements were not driven by chance.

Table 4. Cohen's Kappa for each aggregation approach

	Agreement	Expected Agreement	Kappa	Std.Err.	Z	P-Value
MV	62.86%	43.10%	0.347	0.128	2.71	0.003
SP	80.00%	49.47%	0.604	0.155	3.89	0.000
CW	65.71%	44.16%	0.386	0.133	2.89	0.001
CO	84.85%	50.51%	0.693	0.165	4.19	0.000
AC	97.14%	55.84%	0.935	0.168	5.54	0.000

Finally, the Matthew correlation coefficients (MCC) further confirm the results, indicating the weakest correlation coefficient for the *majority voting* and in contrast, the highest correlation for the *average confidence* approach (Figure 3).

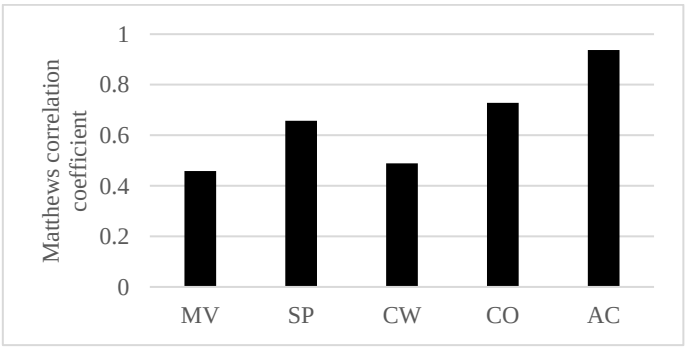

Figure 3. Matthews correlation coefficients

4.2 Further Analyses

We conducted further analyses to investigate how the different approaches perform for questions depending on the level of difficulty of questions. Because all approaches

performed well for easy questions with almost all participants having given the correct answers for this category of questions, we further analyzed the performance of the different approaches for the difficulty levels "medium" and "difficult". When only investigating the performance for the questions in the "medium" category, all approaches perform better compared to the overall performance (Figure 4). In line with the overall performances, the *majority voting* is outperformed and leads to the weakest percentage of correctly identified answers. Further, both the *surprisingly popular* algorithm and the *average confidence* approach provide 100% correct answers.

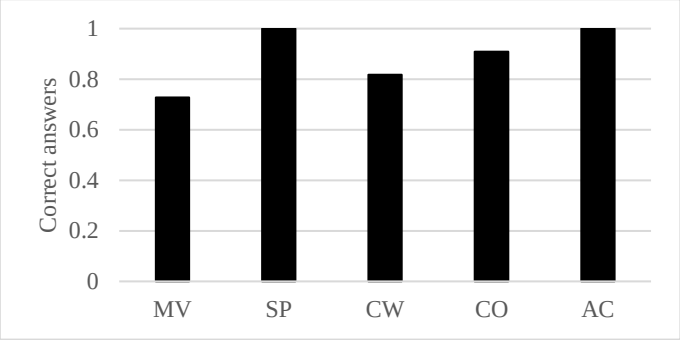

Figure 4. Performance for questions of difficulty category „medium"

Taking the above results into account and comparing them with the overall performances, we assume that the lower overall performances should mainly be driven by the performances for the difficult questions. Hence, we additionally analyze the performance for the most difficult questions.

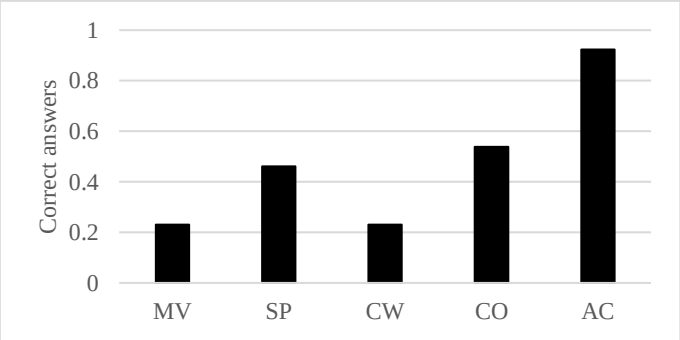

Figure 5. Performance for questions from the category "difficult"

Despite the average confidence approach, all other approaches perform much worse compared to the overall performance (Figure 5). More specifically, the *average confidence* approach provides significantly better results (+38.4%, p= 0.03 by two-sided matches-pair test) than the second best approach, the *confidence only* method. Again, *majority voting* is outperformed by all others approaches analyzed in our study. Hence we can conclude that the average confidence approach leads to the best and most

reliable results, providing the highest percentage of correct answers across all different categories of questions.

5 Conclusion

With the rise of crowdsourcing, the importance of collective decision making has increased and firms increasingly employ mechanisms to aggregate multiple judgements [6]. However, the currently most popular method to aggregate multiple judgements, the *majority voting* approach, is simply aggregating all judgements without taking account of the confidence or expertise of participants. Since this approach carries several limitations and tends to lack reliability [8], we investigated a number of other approaches to aggregate a large number of judgements and compared their performance with each other. Moreover, we differentiated different categories of difficulties for the questions asked and investigated the performances of the different aggregation methods depending on question difficulty. Results indicate that the *average confidence* approach provides the best results across all different categories of questions while the *majority voting* provides the lowest percentage of correctly identified right answers. In contrast to previous literature [4], the *surprisingly popular* approach did not provide the highest percentages of correctly identified right answers in our study. Explanations for this result can be twofold: Compared to the offline experiments conducted in previous research [4], we used a much more heterogeneous and anonymous crowd. In contrast to the crowd used in an offline context, e.g. when asking dermatologists to answer several questions, the crowd used in our experiment had no information about other participants, for example about their educational background. Hence, the crowd was not able to predict the popularity of the answer across the whole crowd correctly since the participants had no information about the level of knowledge of other participants. Second, for the same reason of being unable to estimate the knowledge level of other members in the crowd, people might overestimate the collective intelligence, also leading to wrong results. However, in contrast to the confidence based approaches in our study (CW, CO and AC), the *surprisingly popular* approach needs approximately fitting predictions of the popularity of answers which seems hard to guess for an anonymous crowd.

Our study contributes to literature in several ways. First, we contribute to the literature on collective intelligence by comparing different approaches to aggregating multiple judgements. Our results confirm existing concerns regarding the *majority voting* approach and reinforce the need to develop other, more reliable methods for collective decision making. Second, in contrast to previous studies [4], we use a heterogeneous online crowd on a commercial crowdsourcing platform. By doing so, we apply a very practical approach since not every company has the opportunity to recruit its own internal crowd to answer their questions. Using external crowdsourcing platforms with millions of potential contributors offers a valuable alternative for companies to use almost unlimited personnel resources. The difference in relative performance for an online setting compared to the previously examined offline settings might suggest that

different aggregation methods could differ in performance based on the composition of the crowd. Apart from contributing to literature, our results also have managerial implications for companies currently using or planning to use crowd voting for collective intelligence. Companies should avoid using a simple majority voting approach since it may not lead to reliable results. Accordingly, companies should employ approaches that take an additional input into account, for example, by applying confidence based methods which offer the advantage of including participants' confidence or expertise in their judgements. Finally, when aggregating results companies should consider the origination of the crowd, since it could influence the reliability of the chosen aggregation method. While our study provides important insights relevant for research and practice, we acknowledge certain limitations that ought to be considered. First, the questions used in our experiment are very specific, in order to see if the results are generalizable, further research is needed. Nevertheless, we suggest to carry out additional studies with other types of crowd voting tasks to investigate the influence of the composition of the crowd on different aggregation methods. Second, due to the online setting of the experiment, we cannot rule out that some of the participants look up answers on the internet. However, the design of our experiment allows us to track this behavior to some extent, by counting how often the window with the questionnaire got sent to the background. There seems to be no systematic differences between the different blocks. We hope that our work will open up new avenues for future research, investigating new ways to aggregate multiple judgements and extract the wisdom from the crowd.

References

1. Mack, T. and Landau, C.: Winners, Losers, and Deniers: Self-selection in Crowd Innovation Contests and the Roles of Motivation, Creativity, and Skills. Journal of Engineering and Technology Management 37, 52-64 (2015)
2. Howe, J.: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Business, New York (2008).
3. Chiu, C.M., Liang, T.P., and Turban, E.: What can Crowdsourcing do for Decision Support? Decision Support Systems 65, 40-49 (2014)
4. Prelec, D., Seung, H.S., and McCoy, J.: A Solution to the Single-Question Crowd Wisdom Problem. Nature 541, 532-541 (2017)
5. Bonabeau, E.: Decisions 2.0: The Power of Collective Intelligence. Sloan Management Review 50, 45-52 (2009)
6. Soukhoroukova, A., Spann, M., Skiera, B.: Sourcing, Filtering, and Evaluating New Product Ideas: An Empirical Exploration of the Performance of Idea Markets. Journal of Product Innovation Management (29:1), 100-112 (2012)
7. Hong , L., and Page, S.E.: Groups of Diverse Problem Solvers can Outperform Groups of High-Ability Problem Solvers. Proceedings of the National Academy of Science (101:46), 16385-16389 (2004)
8. Larrick, R.P., Mannes A.E. and Soll, J.B. The Social Psychology of the Wisdom of Crowds. Krueger, J.I., ed. Frontiers of Social Psychology: Social Judgment and Decision Making, 227-242 (2011)

9. Bettman, J.R., Luce, M.F., and Payne, J.W.: Constructive Consumer Choice Processes. Journal of Consumer Research (25:2), 187-217 (1998)
10. Gilovich, T., Griffin, D. and Kahneman, D.: Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, Cambridge, UK. (2002)
11. Simmons J., Nelson L.D., Galak, J. and Frederick, S.: Intuitive Biases in Choice vs. Estimation: Implications for the Wisdom of Crowds. Journal of Consumer Research (38:1), 1–15 (2011)
12. Galton, F.: Vox populi. Nature (75:1949), 450-451 (1907)
13. Lorenz, J., Rauhut, H., Schweitzer, F. and Helbing, D.: How Social Influence can undermine the Wisdom of Crowd Effect. Proceedings of the National Academy of Science (USA) (108:22), 9020–9025 (2011)
14. Keuschnigg, M. and Ganser, C.: Crowd Wisdom Relies on Agents' Ability in Small Groups with a Voting Aggregation Rule. Management Science (63:3), 818-828 (2017)
15. Budescu, D.V. and Chen, E.: Identifying Expertise to Extract the Wisodm of Crowds. Management Science (61:2), 267-280 (2015)
16. Surowiecki, J.: The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. Doubleday Books, New York (2005)
17. Duan, W., Gu, B. and Whinston, A. B.: Informational Cascades and Software Adoption on the Internet: An Empirical Investigation. MIS Quarterly (33:1), 23-48 (2009)
18. Lorenz, J., Rauhut, H., Schweitzer, F. and Helbing, D.: How Social Influence can undermine the Wisdom of Crowd Effect. Proceedings of the National Academy of Science (USA) (108:22), 9020–9025 (2011)
19. Eysenck, H. J.: The Validity of Judgments as a Function of Number of Judges. Journal of Experimental Psychology 25, 650–654 (1939)
20. Preston, M. G.: Note on the Reliability and Validity of the Group Judgment. Journal of Experimental Psychology 22 462–471 (1938)
21. Chen, K., Fine, L., and Huberman, B.: Eliminating Public Knowledge Biases in Information-Aggregation Mechanisms. Management Science (50:7), 983–994 (2004)
22. Aydin, B.I., Yilmaz, Y. S., Li, Y., Li, Q., Gao, J., and Demirbas, M.: Crowdsourcing for Multiple-Choice Question Answering. Proceedings of the Twenty-Sixth Annual Conference on Innovative Applications of Artificial Intelligence, 2946-2953 (2014)
23. Koriat, A.: Subjective Confidence in Ones's Answers: The Consensuality Principle. Journal of Experimental Psychology (34:4), 945-959 (2008)
24. Zizzo, D.J.: Experimenter Demand Effects in Economic Experiments. Experimental Economics (13:1), 75-98 (2010)
25. Mann, H. B., and Whitney, D. R.: On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other." Annals of Mathematical Statistics 18, 50–60 (1947)
26. Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics 1, 80–83 (1945)
27. Cohen, J.: Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. Psychological Bulletin (70:4), 213-220 (1968)
28. Matthews, B. W.: Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. (405:2): 442–451 (1975)