universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„The effects of scale on wisdom of the crowd“

verfasst von / submitted by

Martin Maschl, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien / Vienna 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 915
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Betriebswirtschaft UG2002
Betreut von / Supervisor:	Ass.-Prof. Steffen Keck, PhD

Contents

List of tables

List of figures

Introduction

Idea of wisdom of the crowd came up through an experiment done in 1906, where 800 random participants guessed the weight of an ox. The median was just 1% off the true value. Francis Galton, a British elitist exposed these phenomena and since then many researchers try to use wisdom of the crowd to solve problems.[1]

The term wisdom of the crowd is widely recognized, even if people may not be able to come up with a definition they can still work with the term. Nowadays wisdom of the crowd is not only a theoretical concept to use knowledge or wisdom of many individuals combined to gain estimates that are even more accurate than using knowledge from a single person. The next part shows, that it is a practical method now, which is used for many different aspects and fields in science. Using technologies helps to gather, analyse and evaluate data from many people in a short period of time. Today's technology can be used in many ways like gathering, analysing and interpreting information. Wisdom of the crowd is based on information, knowledge or wisdom from people and then can be combined. Therefore, considering factors, which are or could influence wisdom of the crowd need to be respected, researched and analysed. Therefore, not only controlling known influence factors of wisdom of the crowd is important. Furthermore, contacting research for yet unknown factors should also be done to avoid bias or in worst-case scenario a wrongly interpreted wisdom of the crowd.

To help with such an excessive task this work is split into two main parts. Focus of the first part is to give an overview of research related to wisdom of the crowd and found influence factors as well as the experiments done in those articles. Over these past years many kinds of experiments have been done to identify such influence factors and their effect on wisdom of the crowd. So, the goal was to create a table of research with experiments to list some known influence factors. Second part will be introducing the hypothesis of a new influence factor currently not research. Consequently, an experiment was done, and analysis of the results will be presented and discussed.

[1] (Chowdhury, Burt, Akkaoui, & Davies, 2015, S. 1)

Ideally, in the end this paper should give a quick overview of some already know influence factors on wisdom of the crowd and show persuade the research for yet unknown factors. This should help to either avoid or plan accordingly to influence factors to help and improve future studies of wisdom of the crowd. If possible, this paper can be used as a guide to avoid or prevent biases results by either eliminating or controlling known influence factors for researching additional possible areas of operations for wisdom of the crowd.

Overview of factors influencing wisdom of the crowd

First goal of this work was to create an overview over experiments already done with the focus wisdom of the crowd and/or researched influence factors. Because of that, the first point in this chapter is to introduce experiments to wisdom of the crowd and their known influence factors. The table should provide a table to give a brief overview of the experiment and their found results or effects. To offer a better understanding of such experiments the first article got a detailed summary of the used method and their experiments. Afterwards only a short summary of additional information from each source is given.

Table of related research with similar experiments of wisdom of the crowd

NR.	Articles	*Journal*	Experiments	Results & Effects
1	Intuitions About Combining Opinions: Misappreciation of the Averaging Principle	*Management Science*	In experiment 1&2 choosing a strategy for forecasting Experiment 3&4 made adaptations to limit numerical influence on the forecasting	Misunderstanding of averaging Incomplete information Own imagination and numerical Influence
2	A solution to the single-question crowd wisdom	*Nature*	Collecting people's beliefs about US state and capital questions	Algorithm surprisingly popular Limitations of participants or people within the crowd with their capabilities as well as the available information
3	Flexible Human Collective Wisdom	*Journal of Experimental Psychology*	Participants had to give confidence levels as individual and within groups to decide on 50/50 situations	Group performance outclasses individual ones Following the lead of majority after receiving additional information or having group discussions

				Minority guess only gets picked if they have high confidence
4	How social influence can undermine the wisdom of crowd effect	*Proceedings of the National Academy of Sciences*	Participants answered knowledge questions multiple times within different conditions (aggregated, full and no information)	Three effects are found with influence the wisdom of the crowd Social influence Range reduction Confidence
5	Are we wise about the wisdom of the Crowds? The use of group judgments in belief revision	*Management Science*	Validation of judgments made by different sized groups on 7-point Likert scale; Estimating average high temperature with different sized group sources	Participants rated larger groups as more accurate than smaller groups; Influence of groups increased with size of the group; Performance improvement is superior if influenced by bigger groups than smaller groups
6	Quanty: An online game for eliciting the wisdom of the crowd	*Computers in Human Behaviour*	Participants should estimate different properties (height, weight, depth, temperature, etc.) from pictures	Reliability of the mean is higher compared to the average and standard deviation, because the influence of spikes gets reduced
7	The Wisdom of the Select Crowds	*Journal of Personality and Social Psychology*	Simulation of 3 experiments; Crowds of judges create a guess followed by calculating the absolute error; Expertise calculated by averaging the accuracy over guesses minus the mean absolute error	Select-crowd strategy is precise, vital and more attractive than the normal wisdom of the crowd; Select crowd can be achieved by relying on the top 5 judges within crowds; Select-crowds outperform the whole-crowd and best-member strategy
8	The wisdom of the crowd playing The Price Is Right	*Memory & Cognition*	4 Participants were guessing the prize of a product; Guesses were giving in sequential order with every player knowing the previous guess	Players use some kind of aggregation, either directly on the guesses or with the help of a decision model. The best methods use aggregation compared to the worst methods, which do not use aggregation. Decision models do better than methods if one or more players use some kind of strategy for their guess

9	Wisdom of the crowd and natural resource management	*Trends in Ecology & Evolution*	Using local knowledge to assume natural resource amounts	No correlation between exploited resource to wisdom of the crowd; Inappropriate use of wisdom of the crowd; More studies needed
10	Harnessing the Wisdom of the crowds in Wikipedia: Quality through coordination	*Trends in cognitive sciences*	Combining knowledge to generate articles for Wikipedia	Coordination costs, Quality of the article and number of authors is problematic
11	Investigating the Wisdom of the Crowds at scale	*Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*	The experiment consisted out of multiple choice or point estimations in the areas of explicit, tacit and prediction knowledge	For question level performance the outperformance of wisdom of the crowd can depend on the questions; Category level performance is better than question level performance
12	Matching mechanisms to situations through the wisdom of the crowd	*Thirtieth International Conference on Information Systems*	45 situations which had to be matched with mechanisms and the categorized; Level of participants ranged from novices to experts with 5 years' experience; One expert with 20 years´ experience used for analysis as source	Crowd displayed expertise; Aggregation of matching combines the knowledge and reduces common mistakes; Students which named their categories did better;
13	Rating scales for collective intelligence in innvoation communities: Why quick and easy decision making does not get it right	*Thirty First International Conference on Information Systems*	Web-based experiment to rate ideas; Participants rated those ideas via computer from at home/work/lab;	Simple rating mechanisms (good/bad or 5 star ratings) do not provide effective rankings; Multi-attribute scales outperform simple rating mechanisms considerably
14	Strategies for revising judgment: How (and how well) people use others´opinions	*Journal of Experimental Psychology: Learning, Memory, and Cognition*	Estimating mean annual salaries of schools through; Before the experiment participants guessed: During the experiment participants received randomly answers from another participant and had the option to adjust their own answers;	Averaging proved to be the most effective method; Choosing own expertise for questions about the own country and averaging would create better performance for questions about other countries;

			Meaning advice from participants in the exact same situation was given and received;	Bracketing was not favoured in experiment four; Results show that choosing was the better strategy
15	The wisdom of the inner crowd in three large natural experiments	*Nature human behaviour*	Guessing the amount of pearls in a large transparent container (shape of a champagne glass); Comparing the wisdom of the crowd to multiple individual attempts;	Alternative to a situation without Lower effectiveness than wisdom of the crowd; Time effect between guesses; Amount effect of guesses; Trade-off between time and amount;

Table 1 List of research of wisdom of the crowd with experiments and results

1)

People tend to think incorrectly about averaging. They either use information from others, ignore it or rarely accept it totally without thinking about errors. The average principle uses the average of estimates from different persons to create a better prediction. Thinking about a simple guessing game with the question "how many coins are in the glass?" two people guess 60 and 70 pieces. The correct answer of 73 allows calculating the miss by 13 and 3. Therefore, the average miss of these two people was 8. Now using the average of both guesses gives us 65 which also miss by 8. Now the average principle should always lead to an equal or even better prediction than the average assumption of the crowd. Altering the example could lead to the following, the two guesses are now 60 and 80 which leads to errors of 13 and 7. Their average error changed to 10, on the other hand averaging of the predictions leads to 70, which only has an error of three and outperforms both individual guesses.

This paper contains four different experiments about people´s beliefs and the effect of averaging estimation.

Experiment 1

Participants received a scenario of two analysts and their forecasting patterns with their errors of over-/underestimates. Then participants had to choose one of five strategies.

- *Strategy 1: Retain Ms. A as the dollar/euro forecaster and reassign Ms. B.*
- *Strategy 2: Use Ms. A`s forecast 60% of the time and Ms. B´s forecast 40% if the time.*

- *Strategy 3: Average the two forecasts and use this average as the bank´s forecast.*
- *Strategy 4: Ask Ms. A and Ms. B how confident they are for each forecast. Use the one who is more confident. If they are tied on confidence, go with Ms. A.*
- *Strategy 5: Have Ms. A and Ms. B sit down and discuss their opinions. Require them to agree on a single forecast.* [2]

Looking at the results more than half of the Participants (57%) thought that averaging would be no better than an average judge. On the other hand, most of these 43% thought, that averaging would considerably outperform the judges. Considering all conditions, the majority failed to see the positivity of averaging.

Experiment 2

Participants received a scenario of two managers predicting the number of customers in their movie theatre for forecasting the necessary personal numbers. The difference between the second to the first experiment were adding simultaneous and sequential formats for the judge guesses. Averaging should be recognized within the synchronized format, while the consecutive format should lead to the misunderstanding of the averaging principle. The results were within the expectations and show more misinterpretation of the averaging principle in the consecutive set-up than in the synchronized set-up.

- *Strategy 1: If Ty´s estimate alone were always used, how much money would the theatre have lost per day on average? (Ty alone)*
- *Strategy 2: If Chris´s estimate alone were always used, how much money would the theatre have lost per day on average? (Chris alone)*
- *Strategy 3: If the mean (that is, the midpoint) of Ty and Chris estimates were always used, how much money would the theatre have lost per day on average? (Averaging)* [3]

Results show similar results to the first experiment, in which participants who choose averaging thought of it as the better option to both judges. Also, similar most of the participants supposed that averaging would not perform better than an average judge.

[2] (Larrick & Soll, 2006, S. 115)
[3] (Larrick & Soll, 2006, S. 117)

Experiment 3

Within experiment 3 adjustments were made to test participants understanding of averaging. Therefore, both judges were defined as equally good compared to the right answer. A third judge was included which would always guess the average of both judges. The idea in experiment 3 was to limit numerical information to access participants' knowledge about averaging. Now participants had to decide the average would be better, worse or as good as both individual estimates.

This time result show that 50% of the participants thought that averaging would outperform both judges, 32% thought it would be as good as both individual estimates and only 13% believed it that the judges would be better than the average. Compared to the other two experiments participants were more right about the averaging.

Experiment 4

Participants received a scenario with students running a film series, within their project they had to predict visitors at each movie. Numbers were provided which stated the lowest and highest possible attendance. Additional information about estimates were given and then participants had to choose their confidence about different mottos. Examples for those mottos are:

- *Two heads are better than one*
- *The truth lies in the middle*
- *Compromise leads to mediocrity* [4]

Results show that only a few people answered that averaging would only lead to an average result. The majority of those who misunderstood averaging thought that everything was possible, meaning that averaging could lead to better, even or worse result in the end.

Many diverse factors can influence the misinterpretation of averaging. If people, try to use reason they tend to have an incomplete understanding and depend on their own imagination. This can lead to

[4] (Larrick & Soll, 2006, S. 121)

numbers outside of the scale which can be seen in the data set of the experiment done. Based on that, it is also listed as the first influence factor in the following table.

The following table shows related research to wisdom of the crowd, as well gives a brief overview of their contacted experiments. Below the table each listed article is shortly summarized with more details to the experiments. Furthermore, the results of the research and their influence factors are summarized and listed.[5]

2)

Harnessing knowledge by using the wisdom of the crowd is widely accepted and could be prioritized over consulting expert advice. However, wisdom of the crowd comes with limitations and biases information of the individuals with limits to the range of their focused knowledge. Therefore, instead of a democratic vote which picks the most widespread answer an alternative is needed.

This article tested a method for questions with binary answer-options to show that it is the best option to collect the wisdom of the crowd, rather than democratic voting or evaluating the confidence levels of the answers. This substitute uses the prediction of the distribution of other participants believes and picks the answer that based on this distribution has more votes than predicted. This paper names the algorithm as surprisingly popular.

Different studies were used to examine this algorithm with study 1 using US state capital questions, study 2 general knowledge questions, study 3 diagnose skin images from professional dermatologists and study 4 guessing the value of art by ordinary people and professionals.

Relevance of the surprisingly popular algorithm can be found within markets where individuals trade contracts which are connected to detailed upcoming events. With this algorithm experts can choose an alternative to the majority vote or belief over effects predicted by the wisdom of the crowd. Theoretical situations set no limitations to participant's beliefs and under such circumstance, this algorithm can outperform other voting methods. On the other hand, can be seen within the results that even the

[5] (Larrick & Soll, 2006)

algorithm surprisingly popular will hit its limits with available information to the participants and as well their capabilities.[6]

3)

Based on given information, people make decisions and gaining additional information can lead to a change in the decision-making process. Through wisdom of the crowd, people's knowledge can be combined, and the averaging principle shows that this combined knowledge is more likely to have higher accuracy than that one from one individual. This paper tests if people adapt their decisions within groups because of additional or changed information and input from the group.

Students at the University of California participated in the experiments. Split into thirty participants for experiment 1, which had ten groups with three members and the same spreading concept for experiment 2. In experiment 1 participants had to respond with yes or no if they detected a signal and received the introduction that the signal would be present 50% of the time. In experiment 2 participants had to guess if a sandstorm could happen based on generated environmental data, consisting for example of wind-speed and temperature.

Results show a significant better performance of groups over individuals. Even though this is only the case for the mixture situations within the second half of both experiments. Furthermore, the adaption to additional information lead to selecting the majority guess, except if an individual had a high confidence guess which resulted in selecting the minority guess.[7]

4)

Like the previous papers this one also agrees on the fact, that depending on the Wisdom of the crowd, average judgment can be better than the judgment of an individual. While groups can outperform individuals, their combined knowledge will be influenced by each other. This paper found and name three factors which are generated within groups and influence the wisdom of the crowd.

[6] (Prelec, Seung, & Mccoy, 2017)
[7] (Juni, Eckstein, & Enns, 2015)

The experiment was contacted in a laboratory environment were each participant entered the guess via a separate PC. There was no communication between participants, since all necessary information about the experiment was explained at the beginning and additional in a print version available.

Results show that three factors called the social influence, range reduction and the confidence effect influence wisdom of the crowd. Social influence effect will reduce the variety without reducing the collective error, meaning groups tend to approach other guesses which holds no improvement. Range reduction effect describes the fact that the right answer won´t be found in the middle of the groups guesses. Meaning that groups have a reduced amount of consistency after experiencing social influence. Confidence effect shows that participants gain confidence with additional information, which can be seen in the two condition with aggregated information and full information compared to the condition with no information where participants had fewer confidence levels.[8]

5)

People tend to use advice to adjust their own beliefs or thoughts. Research revolves people are significantly more influence from groups rather than individuals to adapt their point of view. Studies show that people can improve their judgment with the advice of others. Especially if wisdom of the crowd is used and the crowd size increases, people tend to readjust their own beliefs more to the group judgment which leads to a better performance. On the other hand, people will correct their initial belief only slightly if the advice comes from an individual or a small group.[9]

6)

Participants were split into two groups with Group 1 consisted of 46 Americans participants and group 2 consisted of 50 participants from 23 different countries using the metric system. Statistical analysis was split into four parts. First analysing the guesses of each group to calculate averages and comparing the results of each group to each other. Secondly, removing guesses outside the range of the average plus two times the standard deviations and a new average was calculated. Therefore, the results of group 1 improved from 4.3 times higher to 3.08 and group 2 improved from 4.55 higher to only 1.59 times.

[8] (Lorenz, Rauhut, Schweitzer, & Helbing, 2011)
[9] (Mannes A. E., 2009)

Within the third analysis the median was used to calculate the accuracy of estimates. Method four used a transformed median to calculate the accuracy of the participants. Calculations with method 3 and 4 lead to better results of accuracy than methods 1 and 2. Therefore, it can be said that the median is more reliable than the mean and the standard deviations, if there are outliers in the data set.[10]

7)

Alternative strategy called select-crowd uses a ranking system bases on chosen ability. The article experiments with three different strategies. First the best-member, second the whole-crowd and last the select-crowd strategy. These experiments were contacted in a simulated environment with either high or low bracketing combined with either high or low dispersion in expertise creating for classes. For each class a possible best solution strategy was defined.

Select crowds will be more attractive to people who want to include their own believes into generating judgment with help of crowds. Therefore, thus people will either choose the best-member strategy or use select crowds with the top five judges to take a guess. This results from the fact that people can identify the top judges within a group, but are uncertain which individual is the best one. [11]

8)

Data sets consist out of 72 try-outs of the U.S. game show The Price is Right. 11 different methods were used to aggregate the predictions of the players. A summary of all methods is listed in a table, for example using the first guess, average, mean, and average of the two centred guesses or additionally four decisions models. The four decision models ranging from simple models choosing a random number within the boundaries to players having different ranges around the true value. As well as having models, ranging with strategical bids from one player based on the knowledge of player and using the previous bids to calculate possible winning guesses to additional players placing bids with strategy.[12]

[10] (Chowdhury, Burt, Akkaoui, & Davies, 2015)
[11] (Mannes, Soll, Larrick, Simpson, & Kawakami, 2014)
[12] (Lee, Zhang, & Shi, 2011)

9)

Situations with lack of data or even no data at all must rely on local knowledge to assume current or future resource sizes. Especially natural resource management, for example fisheries, use such an approach to avoid too much consumption or overuse of resources. At this point wisdom of the crowd can be used not only to access local knowledge furthermore to gain better results of estimates. With this approach, the following four criteria form a cycle of repetition. Assessment of current biomass, second analyses of reference points, third potential management response and fourth information as well as enforcement. [13]

10)

The use of the internet made access to information from many different people quite easy. Wikipedia is a prime example to access wisdom of the crowd. Although there is need for some coordination to write articles on the website, the quality of those articles is quite similar to traditional encyclopaedias. Coordination can use explicit, which stands for direct contact between the authors or implicit coordination based on structure and models. Research focus was to link the number of authors to the quality of an article. Only a part of the idea wisdom of the crowd applies in this research.[14]

11)

The experiment consisted out of multiple choice or point estimations in the areas of explicit, tacit and prediction knowledge. Fifty categories with 20 questions each were crowdsourced to a group of volunteers. The performance was calculated on two different levels, first the question tier and second the category tier. Generating the median for open-ended tasks and the most popular answer for categorical tasks. Furthermore, adding the effect of social influence on the performance the crowd wisdom to the analysis.[15]

[13] (Arlinghaus & Krause, 2013)
[14] (Herzog & Hertwig, Harnessing the wisdom of the inner crowd, 2014)
[15] (Shankar Mysore, 2015)

12)

The expert assessed all matched answers as practical. Comparing expert and crowd choices resulted into a match of 73,33% for the answers. Analysing all the individuals guesses with second and more choices resulted in a match of 82,22%. The other eight solutions differed heavily between expert and crowd. Another two experts agreed to first expert choices with 7 out of 8 and 6 out of 8. While both other experts agreed once with the crowd, one of them had a solution that did not match any other matches.[16]

13)

Participants had to rate each idea once and were unable to see other people's rating to avoid bias. The task to rate ideas based on their quality, which was split into novelty, relevance, feasibility and elaborateness using the personal experience. Three different scales were used to rate the idea. First one was binary with promote or demote idea and second one was a 5- star rating scale. Last one was a complex rating scale with four 5-point ratings for each split of quality used above. An independent expert rating by a qualified jury was used as comparison to the ratings of participants. [17]

14)

This article researches 2 straightforward response of people in response to gaining additional information from others. First response is to change one´s answer to match those of the other individual and second response is to average the information generating the answer. Four experiments were conducted, additionally to the method of experiment one, in experiment two participants had to rate their familiarity with the schools. Experiment three created an environment in which averaging should be less practical pick. Participants were paired with two different nationalities and asked questions about their both countries. Experiment four included general knowledge questions only for the USA. Additionally, participants rated their confidence level.[18]

[16] (Nickerson, Zahner, Corter, Tversky, & Yu, 2009)
[17] (Riedl, Blohm, Leimeister, & Krcmar, 2010)
[18] (Soll, 2009)

15)

Some situations might not provide data from many different subjects; therefore, an alternative way is needed to collect wisdom of the crowd and use the average principle. For a case like that, this paper tested the idea of using repeatedly judgment from a single person to generate a dataset. This approach is supported by the research like “Stroop, J. R. Is the judgment of the group better than that of the average member of the group? J. Exp. Psychol. 15, 550–562 (1932)”.

The experiment involved guessing the correct amount of pearls, diamonds or casino chips in a large transparent container. The shape of the container resembled a champagne glass. The data sets included all the estimates from all three years in which the experiment took place and contained multiple guess from the same person.

The result shows a difference in the outcomes by using the arithmetic mean compared to the geometric mean. On the one hand, the true value is constantly overestimated by the arithmetic mean, while on the other hand the geometric mean underestimates during the first two years and only overestimates in year three. Consequently, the factor of time between making the estimations is very important. It was found out that the accuracy improves if there is a delay between the two guesses. For situations with limited time, there is a trade-off between extra guesses and extending time between those guesses. Concluding to the fact that aggregation of individual, multiple guess has a lower accuracy than wisdom of the crowd.[19]

Experiment

Introduction

As presented in the chapter above, there are many kinds of influence factors to consider for using wisdom of the crowd. According to the average principle, combining judges’ estimates will lead to an improvement of their accuracy. For aggregating judgment, the experiment done in this work depends on the use of numeric numbers. Therefore, researching the influence of scale is marked as the focus of this

[19] (van Dolder & van den Assem, 2018)

paper. Giving a short example, thinking of two general knowledge questions, first how long is the Amazonas in km and second how much percentage of the world population are female. Initially both questions seem to be equally hard or easy to answer. Within an experiment, participants would guess and try to get as close as possible to the true value.

Taking a closer look at both questions different scales are used and this could hold a possible influence factor. By comparing these two scales, first we have a scale between 0-100 for percentage and second a kind of open scale for kilometre. Based on simple logic the Amazonas cannot have a negative length therefore, the scale only has positive numbers. On the other hand, the upper limit of kilometres is hard to guess. Reminding everyone the Amazonas is 6.400 km long, an information which can easily be obtain by simple typing "length of the Amazonas" into Google. Based on this, for the following experiment the term open scale relates to scales where the boundaries are unknown or hard to guess.[20]

If participants follow simple logic and use the knowledge to see or set limits to a scale, how do scales influence the aggregation and can influence the accuracy of wisdom of the crowd. The experiments goal is to determine if there is an influence of scale on the estimates of the participants. Furthermore, through analysis of the data results should present what kind of influence scales could have. The experiment was conducted in the following steps.

Constructing an experiment around the idea: influence of scale on wisdom of the crowd

At the start, a choice had to be made in form of which kind of scales should be tested. Therefore, to capture the influence of scale on wisdom of the crowd two different scale types were used. For general knowledge questions, defining either an open scale or a closed scale seems difficult. Therefore, the following definitions will be used for this paper. Closed scales are referring to a scale with clearly defined border, for example percentage, which is widely known to start by 0% and end by 100% without having any negative numbers. Open scales are referring to a scale which has non-quantified borders. In

[20] https://www.google.com/search?client=firefox-b-d&q=lenght+of+the+amazonas

this case, participants can only imagine the limits of the scale. To capture people´s unexercised estimates the experiment is split into three parts.

The first part will use an open scale to collect people's beliefs about some facts without giving advice. To convert the idea of the experiment the second part uses a closed scale so that this data can be used as comparison. This second part plays additionally the role of a puffer between the first and the third part, so participants switch their thoughts to another similar task before receiving additional information for the third part. Furthermore, part two should provide data of general knowledge question with the use of a closed scale.

For the first and third part of the experiment, the scale of kilometre was chosen. Participants had to guess airline – distances between two capital cities. By using the same scale and question model for both the first and third part it is possible to collect and review changes in data from participants after they receive additional information.

Preparation for part one and three were done with the use of Microsoft Excel and the website (https://www.luftlinie.org/). Started with creating a list of all capital cities then using the random function of excel to generate 20 pairs of capital cities. Because excel has no option to save random generated numbers or for this use pairs, it was necessary to copy the results of one random generated run. For those pairs the airline distance was calculated, and their corresponding countries were added to give participants additional information.

Preparing the experiment

One important aspect of the question sheet was to split part one and three so that participants would not see both at the same time. Using print material lead to the idea of two-sided printing. Participants had to complete part one completely on the front page. The first half of the second part was on the front side while the rest of it was located on the backside. Idea of this was to occupy participant's attention on the closed scale before receiving advice for the third parts open scale questions.

To reduce explanation time each the experiment started with a short introduction section. Three main points where established. First that the research question would not be explained before answering all

the questions. Second the composition of the questionnaire with an additionally request to answer all question or give an estimate. Third and last one was the request not to skip any part and answer all the questions of each part.

Part one started with a short explanation about the task, to guess airline-distances between two capital cities. Part two consist of questions out of the article (Herzog & Hertwig, Think Twice and then: combining or choosing in dialectical bootstrapping?, 2013). Followed by part three with a reminder of the explanation of the first part and additional advice. Advice had the form of setting the boundaries to the open scale while giving the smallest and largest distance between two capital cities.

The following table shows an example for the format of the first and third part. As seen in this table additional information was available in form of country name to the capital cities. By using this, format participants only had to enter their assumption in the empty box placed in the middle.

Country	1. City	Distance in km	2. City	Country
France	Paris		Kigali	Rwanda
Kazakhstan	Astana		Ulaanbaatar	Mongolia

Table 2 Form of part one and three

Furthermore, advice was given in the same format, which should be easily understandable for each participant. Size of the table was made a little bit smaller to differ from the part where participants had to fill in their own guesses and a headline was added as can be seen below.

Limits to the scale:

Country	1. City	Distance in km	2. City	Country
Italy	Rom	4,94 km	Vatican	Vatican City
Spain	Madrid	19.861 km	Wellington	New Zealand

Table 3 Advice table with headline

Part two used a different format to be easily differentiated from the other parts shown in table 4. Because part two uses questions that should be answered the introduction was just a simple line to answer the following question. As seen in the example the format should eliminate confusion and the task should be easy to understand and quickly to finish.

Question	Your Guess in %
What percent of the world's airports are in the United States?	
What percent of the earth's surface is covered by water?	

Table 4 Form of part two with questions from (Herzog & Hertwig, Think Twice and then: combining or choosing in dialectical bootstrapping?, 2013)

Procedure of the experiment

The experiment was conducted during a session of Professor Keck course of strategic decision-making at the University of Vienna. It had 52 participants consisting of master students from economics. For this experiment, personal data of participants was not included. The professor provided a time limit at the end the lesson. Participants had enough time to finish the experiment, which can be seen in the first step of data analysis.

Data Digitization

Each questionnaire was set up with a number to enable later examination and safety checks of the answers. This method enabled an easy way to crosscheck each answer with limited workload and reduced time of data digitization. Then all answers were transferred into an excel file for the analysis process. A data form was chosen to be intuitive understandable. First row shows description of each column and thereafter each one of the 52 answers sheets was listed in a single row while each cell was filled with one answer or stayed empty if no answer was given.

Data Analysis

Following the data digitization was the analysis of the data set of the experiment consisting out of 52 responses over three parts with 10 questions per part and was done in Microsoft office excel. An overview of the data was established by splitting the analysis into parts. Starting with an overlook over the data set followed by a deep analysis of each part and the research focus.

Overview

As already mentioned, the data set consists out of 52 responses. The total number of possible answers would have been 1.560, which was not reach due two participants failing to answer all question. The number out guesses recorded was 1.552 out of 1.560 which represents a quote of 99,49%. It is possible

to say that participant 32 missed question 7, while it is safe to say that participant 40 just gave up at the end and did not answer the questions 24-30 at all. The lowest and highest guess for each question is listed in table 5 from the appendix.Once again, the boundaries for Questions 1-10 and 21-30, the min value is 4,94 km and max value is 19.861km. Which leads only to outliers greater than the upper limit and none beneath the lower limit. This is because airline-distances can only consist of positive numbers and participants seemed to understand this very well therefore no negative numbers were recorded. On the other hand, the upper limit was not easy to guess.

A possible way to assume the upper limit could be a mathematical approach. First step could be to simplify the problem by considering earth as a perfect sphere. Now to calculating the circumference gives the length around the world. This leads to the upper limit through logically using only half of it. Because whenever searching for the longest distance between two points on a sphere half of the extent will always be the maximum length. Using this approach would need knowledge of either the radius (6.378,137 km) or the length of the equator (40.075,0167 km) and the formula for the length of circumference.[21] This should then lead to an approximate value of 20.000 km as the maximum value.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Number of guesses for >=20.000	1	1	1	2	3	4	1	1	2	2

Table 5 Number of guesses outside the boundaries of km-scale

Based on this logical approach the following results in table 6 outside of the upper limit were found. As listed, there are very few assumptions of 20.000 or more. After tracking of the numbers, only four participants were identified and those were participants 3, 5, 7 and 28. While participants 3 and 7 had each seven guesses outside of the boundaries, participants 28 had three and participant 5 only had one guess outside of the upper limit.

	Country	City	Distance in km	City	Country
Q5	Austria	Vienna	8.277,15	Seoul	South Korea
Q6	Papua New Guinea	Port Moresby	14.243,75	Vaduz	Liechtenstein

Table 6 Additional information participants received

[21] Exact numbers were found on https://en.wikipedia.org/wiki/Equator

In the first part, Question 5 and 6 were those which true value had the largest distances. The true values of both questions can be seen in the table 7 above. After participants received additional information, the data set showed not a single guess outside of the boundaries for the questions 21 to 30. This fact is also listed in table 5. Here the max values of guesses are all below 20.000. The average of the maximum assumption dropped from 36.600 to only 13.586. On the other hand, the average minimum assumption only slightly changes from 531 to 473. Comparing those spikes to the closed percentage scale showed the following results.

Q-Nr.	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Max	90	85	90	80	81	85	70	90	96	98
Min	2	45	35	1	2	2	0	25	15	16

Table 7 Min and max for the questions with %-scale

Within table 8 can be seen that not a single guess was outside of the boundaries. No information was given to the participants, but all participants were students of the University of Vienna. Therefore, the possibility that every single one knew that percentage range is from 0-100 was extremely high.

This shows the first effect of scales onto the wisdom of the crowd. Open scales can have multiple assumptions outside of the scales range. Otherwise, if a known scale, like percentage, is used the likelihood of spikes will be very small or it is possible that no outliers at all are within the data set. This leads to an impact on calculated numbers for example averages, minimum and maximum values and so on. The exact impact on some of these numbers is going to show up later in the analysis.

A few interesting side facts were overserved during the data digitalization. Some participants misunderstood the number field and filled in their student ID number on experiment sheets. This field was used for reference numbers during the digitalization process. One participant used a shortcut to fill in guesses for the kilometre scale. Instead of writing three zeros and “k” for kilo was used. Lastly, in part two some participants added the percentage symbol (%) after their guess, but not once was the short cut for kilometre (km) added.

Best and worst performer of the experiment

To find the best and worst performers an analysis of the absolute errors (AE) of each question guess was done. For this purpose, a ranking system was established. The smallest AE was placed on the first place.

The second smallest AE was placed second and third likewise. If more than one participant used the same estimate leading to the exact same AE, placement was share among them. This analysis was done for each question and for example, results of question 1 can be seen in table 9 in the appendix.

Likewise, to question 1 most of the other question's rankings resulted in a few ties. Especially question 12, 13 and 15 had many ties. The average ties for each question was around 10. However, for those three questions the ties increased to 30, 28 and 22 accordingly to their listed order. The next highest number of ties was recorded with 14 for question 26. The lowest number of ties was four, for the questions 6 and 23. As explanation for this can be said, that most participants used whole numbers for the percentage-scale. Only 7 out 520 guesses were decimal numbers. This is like the 9 answers of part one and 15 answers of part three which did not end with zero. In addition, 23 assumptions of part one and 19 assumptions of part three did not end with double zero. Meaning that most participants used round numbers for their most of their answers. An example for this is shown in table 9 where the first placement ties between five participants with a guess of 6.000.

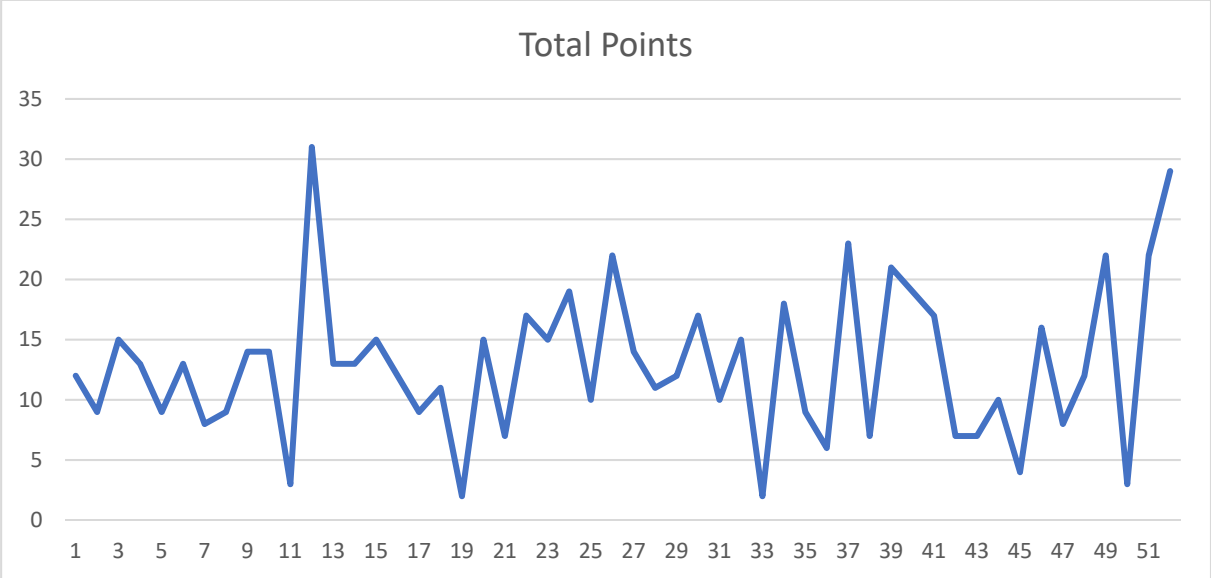

Figure 1 Total Points with points system

Participants scored between 2 and 31 points. This implies that every participant reached at least ones a placement among the top 3. Participant 33 and 19 each scored only two points with the placements of once second and twice third place.

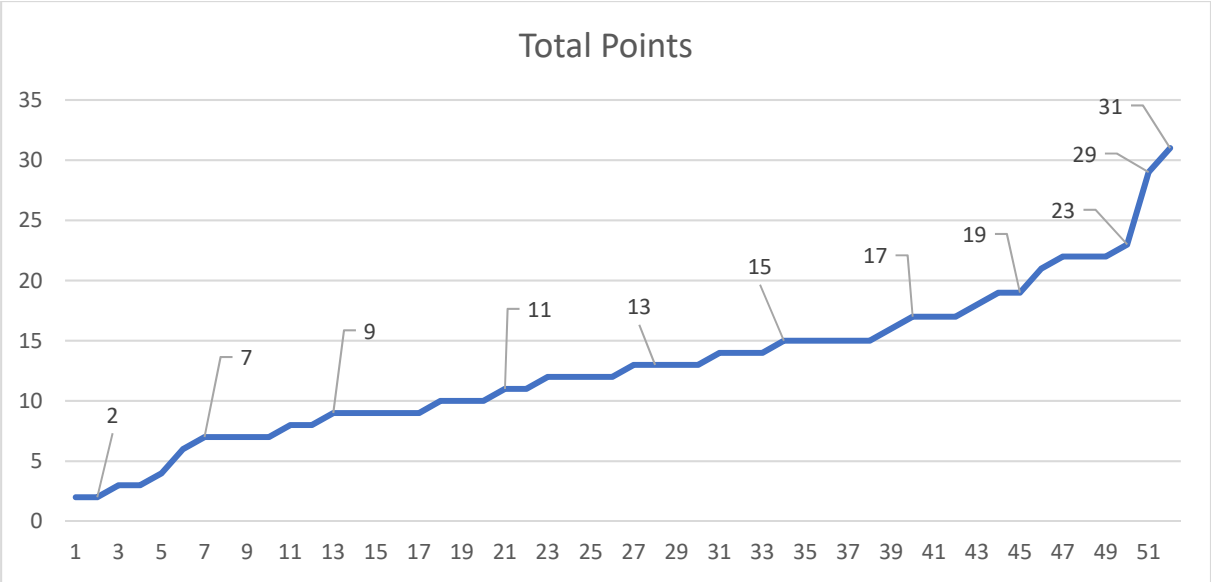

Figure 2 Total Points in rising order

By looking at the total points in an ascending order, a steady rise in points can be seen. Figure 2 draws a good picture of the rise from total points and then the jump near the end from third place to second and first. Besides those jumps at the end, there are only two small jumps between 4 to 6 points and again from 19 to 21. Otherwise, it is a constant increase from 2 to 23 points. The average points were calculated with 12,9.

This is based on the points system, which was used to reward first place with 4 points, second with 2 points and third with 1 point. The ranking is based on the total number of points reached. This was done to establish a difference for the top 3 placements. Therefore, the number of placements alone is not enough to score high total points. The idea follows the popular saying: "quality instead of quantity". Which lead to the scores in the table 10 below.

Player	First	Second	Third	Total number of top 3	Total points
12	5	3	5	13	31
37	2	6	3	11	23
15	1	2	7	10	15
40	2	3	5	10	19
26	4	1	4	9	22
39	3	3	3	9	21
41	1	5	3	9	17
52	6	2	1	9	29

Table 8 Most top three placements

Table 10 shows an order after total number of top 3 placements. With the point system second and third place changed because the quality of the placement was accounted for. Because of that, participant 52 could reach the third place in the end. Winner with the most points was participant 12 (Gold / 31 points), followed by 52 (Silver / 29 points) and 37 (Bronze / 23 points). Although participant 52 is listed at the bottom of table 10, with six first places, the total points were enough to score the second place overall.

Finding the worst player of the experiment was first approached with a similar method. Simply taking the highest number of last places seemed fine at first. There was a surprisingly high gap between the worst five contestants.

Even though contestants 7 and 3 got the most placements at the bottom, their scores with the point system did not rank them at the end. Contestant 7 reached 8 points with two first places and positioned at 48. Contestant 3 reached even more with two first, two second and three third placements getting total points of 15 and positioned at place 15 in the end. Analysing results from the other contestants of table 11 in the appendix also did not support a respectable view to the results. None of these five contestants scored under 4 points. This is listed in table 12 in the appendix.

Because of that, those last placements did not provide a comparable view to the leader board in table 10 in the appendix. Based on this the leader board with the point system was used to find the worst contestants of the experiment.

Now table 13 from the appendix shows the contestants with the lowest number of total points. Comparing those with the contestants placing frequently at the end, it is clear to see that even though guessing completely wrong at some questions does not correlate with total points. Additionally, comparing table 11 and 12, four of five of those participants even reached once or twice the first placement. Because of this, those five participants could outperform those participants in table 13 based on the number of good placements.

Concluding the top 3 players in the experiment where found based on the points system with 4/2/1 points for each of the first/second/third placements. To find the worst performance first the placements at the end where analysed. By comparing those results with the leader board, an unclear view was found.

Players who placed frequently at the end did not have the lowest points. Some of those players scored not bad at all. Therefore, the worst players were found through same system as the best players, ranking through the top 3 placements with the point system.

Part one of the experiment

Short overview of the data set for the first part revealed only one missing guess for question 7 from participant 32. Rest of the set consists out of feasible answers. Meaning at the first look no guesses were found which seem deliberate high or low. Troll answers were also not found, for example negative numbers or answers which were not numerical.

Number of placements in top 3		
	Player	Total Top 3
1	23	5
	39	
	41	
2	1	4
	40	
	46	
3	18	3
	24	
	25	
	30	
	37	

Ranking with point system		
	Player	Total Points
1	39	15
2	18	10
	23	
	30	
3	32	8
	41	
	46	
4	1	7
	24	
	25	
	37	
	40	

Table 9 Comparative table between number of placements and ranking with point system of the first part

Table 14 shows comparison between the rankings of the players with the total number of top 3 placements to the ranking with the point system. Likewise, the analysis in the last chapter this overview shows that participants with high number of top 3 placements are not assured to be at the top of the ranking. In this case only participant 39 with five top 3 placements reached 15 point and gain solo first place for the ranking in the first part. By comparing participant, 41 (marked in green) and 18 (marked in blue), it is clear to see quality matters more than quantity. Here participant 18 reached 10 points, while participant 41 only reached 8 points. This leads to the result of the top placements on the right side in table 14. Participant 39 reached 15 total points with five top 3 placements and scored the first place alone. On second and third place ties between multiple participants accrued.

For each question, the min and max were already calculated and are listed in table 5, as well as the guesses outside of the boundaries for the first part in table 8. As mentioned, outliers were only found within the first part of the experiment. Furthermore, data analysis of the first part was done with attention to the principle wisdom of the crowd. Therefore, individuals' level of judgments were compared to the average level of the crowd. The best guess was specified with the smallest absolute error. The absolute error (AE) calculates through best guess minus true value and is always listed as a positive number.

Table 15 from the appendix shows that four out of the five best answers consisted out of round guesses. Only for question 4, the best answer was not a round number with 3207. Table 16 in the appendix shows the same procedure as 15 with the round numbers. Here only question 10 had the best guess with 2101. As can be seen in table 15 and 16, the best guesses are quite close to the true value. Some participants either had a very good understanding of the true value or got lucky by guessing it nearly perfectly.

Looking at the worst guesses the first five-question show some quite high numbers. Based on the definition for open scale used in this paper those spikes were anticipated. On the other hand, it is a surprising that these numbers had been that high. Compared to the best guesses, table 17 and 18 from the appendix show that without a firm understanding of the wanted answer it can lead to guesses far from the true value. Extreme cases can be seen for question 6, 7 and 9 where the answers are not only higher than half of the earth periphery, which is the possible cap for these questions. It is even higher than the whole earth coverage, which is around 40.000km. In this part the worst guesses match with the maximum values found for all questions of this part.

Looking at the questions to compare the different kind of averaging the wisdom of the crowd, arithmetic mean, geometric mean and median. Many studies have already shown that those methods either outperform or perform on the same level as complex strategies.[22] Now for the questions in this experiment different forms of means were analysed. Comparing median and arithmetic mean shows the advantage of the median in form of less sensitivity to outliers.[23] Often the arithmetic mean overestimates based on such outliers. Here the geometric mean could be better, because it will always be lower or even

[22] (Mannes, Soll, Larrick, Simpson, & Kawakami, 2014, S. 276-277)
[23] (Chowdhury, Burt, Akkaoui, & Davies, 2015, S. 216)

to the arithmetic mean.[24] Advantage of both means would then be that every single guess influences the average with their numerical value and instead of only influencing the position of the median. Table 19 and 20 in the appendix displays not a clear picture of the best averaging solution.

The best performance of the averaging methods for each question is marked green. Arithmetic AE could score the most with four good performances, followed by geometric AE and median with each three times. By comparing these results with the individual AE from each guess, the outperformance of averaging methods can be calculated.

	Q1	Q2	Q3	Q4	Q5
Arithmetic AE	100,00%	63,46%	100%	75,00%	80,77%
Geometric AE	88,46%	96,15%	80,77%	82,69%	96,15%
Median AE	80,77%	94,23%	59,62%	88,46%	94,23%

Table 10 Outperformance of the averaging methods AE´s over the individual AE´s of the first part (1/2)

	Q6	Q7	Q8	Q9	Q10
Arithmetic AE	71,15%	76,92%	86,54%	57,69%	36,54%
Geometric AE	55,77%	53,85%	86,54%	100,00%	98,08%
Median AE	55,77%	50,00%	86,54%	80,77%	98,08%

Table 11 Outperformance of the averaging methods AE´s over the individual AE´s of the first part (2/2)

Now with the outperformance in percentage for the different averaging methods additional details have been found. For example, at question 8 now it can be seen that even though all three methods have different AE`s the percentage of outperformed individuals is even. All additional even outperformances are marked blue. The average outperformance of the averaging methods was calculated with 74,81% for arithmetic AE, 83,85% for geometric AE and 78,85% for median AE. The results show only small differences between the three methods nonetheless geometric averaging shows the best performance overall for the first part of the experiment.

To summarize part one it can be said, that the best guesses came close to the true value, but the worst guesses were far away. Without knowledge or an idea of the limits for the scale, some participants guessed way over the possible airline distances. The average methods were analysed, and results showed that the geometric average had the best AE´s overall. One reason for this can be based on the worst

[24] (van Dolder & van den Assem, 2018, S. 21-22)

guesses and that geometric means are not that sensitive to spikes compared to the arithmetic mean. Even the median could slightly outperform the arithmetic mean in this part.

Part two of the experiment

Short overview of the data set for the second part showed no missing answers. The rest of the set consists out of feasible answers. Meaning at the first look no guesses were found which seem deliberate high or low. Troll answers were also not found, for example negative numbers or answers which had no numerical value.

Number of placements in top 3		
	Player	Total Top 3
1	12	5
	13	5
	26	5
	34	5
	37	5
	52	5
2	3	4
	15	4
	28	4
	31	4
	39	4
	47	4
	48	4
	49	4
3	1	3
	2	3
	4	3
	5	3
	6	3
	9	3
	10	3
	20	3
	27	3
	36	3
	41	3
	44	3
	46	3
	51	3

Ranking with point system		
	Player	Total Points
1	12	15
	52	15
2	49	14
3	13	12
	34	12
4	37	10
	51	10
5	26	9

Table 12 Comparative table between number of placements and ranking with point system of the second part

Table 23 shows comparison between the rankings of the players with the total number of top 3 placements to the ranking with the point system. Similarly, the analysis of part one only relaying on a high number of top 3 placements does not promise a high ranking at the top. In this case, player 49 (marked in green) reached four top 3 placements and could score 14 points which lead to a solo second place. Compared to player 26 (marked in blue) with five top 3 placements could only score 9 points and did not even reach a high ranking. Even though multiple participants had four or five total top 3 placements only two reached 15 points and one 14 points. The number of players with many good placements greatly reduced to a small number in the top 3 ranking. This again supports the use of the points system to differentiate between number of top 3 placements and their value for the overall ranking. Also, for this part, the min and max values of the questions can be seen in table 5. Within part two, no guesses outside of the boundaries of the percentage scale were found. Correspondingly analysis in part one the best and worst guesses are listed in the tables. Table 24 and 25 from the appendix show the best guesses for the second part and only for the last three questions did not hit perfectly. Compared to the first part of the experiment where not a single perfect guess was seen here 70% of the questions were answered perfectly by a least one participant. This shows another effect of the scales in this case of a small range closed scale percentage.

As already mentioned, even the worst guesses are inside the possible boundaries of the scale. Nonetheless, some questions show a high AE, for example questions 1 and 20. The rest displayed in table 26 and 27 in the appendix are AE´s between the value of 26 and 60. This supports the effect that for closed scales it is highly likely to collect data sets without spikes outside of their boundaries. Next step is the analysis of the averaging methods, same principle as in part one. Starting with comparing the best AE results and then the percentage outperformance of these methods matched to the individual answers.

The best option for each question is marked green. In this part, the arithmetic means and median each generate AE´s, which are four times the best option for some question. The geometric mean could only outperform the other two methods twice. Especially for question 12 and 13 are the methods impressive accurate.

	Q11	Q12	Q13	Q14	Q15
Arithmetic AE	71,15%	96,15%	92,31%	57,69%	86,54%
Geometric AE	57,69%	57,69%	92,31%	46,15%	57,69%
Median AE	51,92%	57,69%	92,31%	57,69%	57,69%

Table 13 Outperformance of the averaging methods AE´s over the individual AE´s of the second part (1/2)

	Q16	Q17	Q18	Q19	Q20
Arithmetic AE	42,31%	80,77%	40,38%	71,15%	46,15%
Geometric AE	51,92%	55,77%	32,69%	76,92%	46,15%
Median AE	46,15%	57,69%	42,31%	73,08%	48,08%

Table 14 Outperformance of the averaging methods AE´s over the individual AE´s of the second part (2/2)

Now with the outperformance in percentage for the different averaging methods additional details have been found. For example, at question 13 it can now be seen that even though all three methods have different AE`s the percentage of outperformed individuals is even for all three methods. All additional even outperformances are marked blue. Additionally, can be said to question 13 that the median has zero error to the true value, but in table 30 is only listed with 92,31%. This is because only the outperformed individuals are shown with this percentage number. Meaning that better and even performances are count for the rest of the 100%. In case of question 13 had 7,69% of the participants an even performance to the median.

Now the average outperformance of the averaging methods was calculated with 68,46% for arithmetic AE, 57,50% for geometric AE and 58,46% for median AE. Compared to the first part the outperformance of the averaging methods dropped, also the difference between averaging methods changed. Now the arithmetic mean is the best performing method while geometric mean and median took drops in performance compared to part one. While the spread in the first part was around nine percentage now in the second part, it is a width of around eleven percentage between the best and worst average methods.

Summarizing part two shows that the best guesses came close to the true value and some even hit perfectly with zero error. Not a single outlier was found, a possible reason for this could be the used scale percentage. Because most people or in this case all participants had a very good understanding of the boundaries of the percentage scale. Furthermore, the average methods were analysed, and results showed that the arithmetic average had the best AE´s overall. A possible reason for this could be that the geometric mean did not perform as well as in part one without spikes. Similarly, even though the

median had one a perfect hit, the average performance was just slightly better compared to that of the geometric mean. The best solution for this part is therefore the arithmetic mean even though the performances of all averaging methods in this part was not particularly good compared to part one.

Part three of the experiment

Short overview of the data set for the third part revealed seven missing answers. All the missing data was traced to a single individual. It looks like participant 39 just did not answer the rest of the experiment after question 23. The rest of the set consists out of feasible answers. Meaning at the first look no guesses were found which seem deliberate high or low. Troll answers were also not found, for example negative numbers or answers which were not numerical.

Number of placements in top 3		
	Player	Total Top 3
1	12	6
2	15	4
	20	4
	24	4
	40	4
3	3	3
	10	3
	22	3
	26	3
	37	3
	48	3
	51	3

Ranking with point system		
	Player	Total Points
1	12	13
2	20	11
	24	11
3	22	10
	51	10
4	26	9
	40	9

Table 15 Comparative table between number of placements and ranking with point system of the third part

Table 32 shows comparison between the rankings of the players with the total number of top 3 placements to the ranking with the point system. Similarly, both analysis of part one and two only relaying on a high number of top 3 placements does not guarantee a high ranking at the top. In this case looking at player 15 (marked in green) which had four top 3 placements and did not make it into the ranking of the best four at all. With six point's player 15 only reached rank seven. Player 51 (marked in blue) had only three top 3 placements but could score 10 points due to ranking twice at first place and once at second, which lead to the third place within the ranking. This again supports the use of the points system to differentiate between number of top 3 placements and their value for the overall ranking.

To follow the same pattern as for the last two parts, the min and max values of the questions can be seen in table 5. Within part three, no guesses outside of the boundaries of the scale were found. Like the analysis in the last two parts the best and worst guesses are listed in the tables.

Table 33 and 34 from the appendix show the best guesses of the participants for part three. Most of the answers came close to the true value. Compared to part two not one guess hit perfectly which is similarly to part one. Table 35 and 36 in the appendix display the worst guesses for part three and even though the worst result into high AE´s, they show no extreme highs or outliers compared to part one. Likewise, to part two even for the worst guesses there are none which are outside of the boundaries. Next step is the analysis of the averaging methods, same principle as in part one and two. Starting with comparing the best AE results and then the percentage outperformance of these methods matched to the individual answers.

The best performance for each question is marked green. Arithmetic AE could score the most with five good performances, followed by geometric AE with three times and last the median with only two. By comparing these results with the individual AE from each guess, the outperformance of averaging methods can be calculated.

	Q21	Q22	Q23	Q24	Q25
Arithmetic AE	76,92%	94,23%	100,00%	78,85%	80,77%
Geometric AE	82,69%	61,54%	59,62%	55,77%	94,23%
Median AE	82,69%	78,85%	65,38%	61,54%	80,77%

Table 16 Outperformance of the averaging methods AE´s over the individual AE´s of the third part (1/2)

	Q26	Q27	Q28	Q29	Q30
Arithmetic AE	34,62%	100,00%	38,46%	57,69%	100,00%
Geometric AE	50,00%	86,54%	51,92%	65,38%	86,54%
Median AE	50,00%	73,08%	42,31%	57,69%	67,31%

Table 17 Outperformance of the averaging methods AE´s over the individual AE´s of the third part (2/2)

Now with the outperformance in percentage of the different averaging methods additional details can be seen. For example, for question 21 the geometric AE and median AE have different values but the outperformance over the individuals is equal for both methods. All additional even outperformances are marked blue. Table 39 shows for question 21 an outperformance of 82,69% for the median AE. Same as in part two only the outperformed individuals are shown within the percentage number and the ties at the top lead to shrink the medians AE outperformance to this level instead of being 100%.

Now the average outperformance of the averaging methods was calculated with 76,15% for arithmetic AE, 69,42% for geometric AE and 65,96% for median AE. Compared to the previous parts the outperformance of the averaging methods improved. Arithmetic achieved its best value and for both geometric and median part two could be topped. While the spread was around ten percentage, the arithmetic AE topped the previous highest value by 1,34% and the other two placed between the values of part one and two.

To summarize part three, the best guesses came close to the true value, even the worst guesses were inside the boundaries of the scale. This is because participants got additional information to get a clear picture on the width of the scale for part three. Without any spikes again the arithmetic mean had the best performance overall of the three averaging methods. Median and geometric mean came close, but the arithmetic mean had three perfect performances within the third part. For question 23, 27 and 30 which lead to best performance of the arithmetic mean over all three parts.

Comparing part one and three with the scale of kilometre

Reviewing the results of part one and three will be done with a focus of the wisdom of the crowd and the possible influence of the scale. Main difference between one and three was that participants did not know the boundaries of the scale within the first part and got additional information for the third part consisting of the exact boundaries for the scale of airline distance in kilometre. Therefore, the impact of this difference will now be analysed.

The biggest impact were the missing outliers in part three. Compared to part one with 18 outliers in total, improvement can be clearly seen in part three. Participants, which generated outliers, used the additional information to improve their understanding of the scale and produced better guesses. Looking at participant 3 which had zero top 3 placements in the first part could improve with the additional information in part three and scored top 3 placements three times. On the other hand, participant 7 who had alike participant 3 many outliers in part one could not score any points for the ranking in part three. Despite that guesses, as well as AE improved because outliers weren´t generated at all.

Analysing differences for everyone was done in two steps. First step was to calculate the average answer of every person for the questions of each part. Second step was to compare these results to part three

and look at the differences. Table 41 in the appendix shows that more individuals lean towards lower guesses than before. Here we also have those individuals involved, which guessed outside of the boundaries and could come up with answer inside of the scales borders after receiving additional information. Even though this high numbers were, reduce drastically the average difference only lies around minus 3423. Therefore, it is like the plus 3329, which were added by the other participants to correct their guesses. The total value of those 30 participants, which guessed lower, resulted in a difference of minus 102.700. The total positive difference got up to around 69.910 and resulted in an overall difference of minus 32.790. Comparison between the following questions that have similar true values was done. Analysing the differences between given guesses, min, max and averaging methods. Table 42 from the appendix shows the questions of part one parallel to questions of part three with quit similar true values. Purpose of this analysis is to identify if there was a difference of participants approach or answers after getting additional information.

Ranking with point system					
	Player	Total Points		Player	Total Points
1	39	15	1	12	13
2	18	10	2	20	11
	23			24	
	30		3	22	10
3	32	8		51	

Table 18 Compared ranking of first part to third part

As can be seen in table 43 the best players changed from part one to part three. Interesting fact that not even one could reach the top three ranking in both parts. Looking at both top players, 39 scored first place in part one and unfortunately reached zero points in part three. Player 12 scored first place in part three and reached three points in part one. There is no clear evidence that knowing the boundaries will improve the individual judgment for similar tasks. As the example shows for the two top players, it can result in rising the ranks or dropping down. This holds up for the rest of the participants, some improved and others were less accurate in part three.

	Arithmetic mean	Geometric mean	Median		Arithmetic mean	Geometric mean	Median
Q5	9.525,00	7.985,17	8.900,00	Q22	7.542,71	6.001,11	7.000,00
Q8	7.430,60	5.942,09	6.000,00	Q29	9.442,18	8.531,56	9.000,00
Q9	8.166,38	6.440,91	7.100,00	Q27	6.160,82	5.013,56	5.000,00

Table 19 Comparison of averaging methods part one to three

Table 44 shows a decrease in both the arithmetic and geometric mean for two questions pairs. Additionally, the median decreased once, but otherwise increased for the other question pair. On the other hand, the pair of question 7 and 27 shows a decrease for all three averaging methods.

Geometric mean and the median showed improvements within this analysis of AE and the average AE over part one compared to part three. Looking at all the differences from part one, it displays eight questions below the true value and two above true value for the arithmetic mean. These differences were calculated by subtracting the true value from average mean. Analysing difference showed that values below are smaller than values above true value. This generates an average of -46,90 for the arithmetic mean. For the geometric mean, it was the opposite with eight times above true value and twice below, resulting in an average of 1.613,65. The median showed similar results like the arithmetic mean. Here also eight times below true value and twice above in part one.

Part three generated results for the arithmetic mean with seven times lower than the true value and three times above. The average of those resulted in -591,70 which shows a reduction to part one. Also, the numbers from the geometric mean show a huge reduction with seven times higher and three times lower than the true value resulting in an average of 408,28. For the median an average of -59,91 was found in part three. Table 45 and 46 from the appendix shows an improvement for arithmetic mean twice and once a higher AE resulting in a decrease in performance. Both the AE´s of geometric mean and median resulted in higher errors for part three. Also, for the percentage, outperformance geometric mean and median could not improve in the third part. Only the arithmetic mean AE could improve twice and had once a drop in the performance to only around 57,69% for question 29.

Summarizing the comparison of those two parts resulted in both better and worse performances in part three. Comparing questions with similar true values of part one to part three has shown no clear evidence on how the performance can be improved due additional information of the scale. Some participants could improve their guesses others were better of not knowing the exact borders of the scale. One thing for sure was those outliers can be eliminated with the use of know borders of a scale. Without outliers, the performance of geometric mean and median dropped below the performance of the arithmetic mean.

This shows that the arithmetic mean is the best option for wisdom of the crowd within closed scales. Through the additional information with the borders of the scale, the ranking showed completely different players at the top.

Analysis individuals and wisdom of the crowd

The following analysis has the focus on wisdom of the crowds. A section will be the comparison of the averaging methods with the individuals. Comparing those between questions of the first part and three part. Starting with the percentage errors of the individuals for each question. Calculated by guess divided by true value multiplied with 100. The average of first part and third part were compared to look for improvements. Looking at all participants, 29 out of 52 could improve their judgment. The range of the differences starts by -128,12 and goes up to 405,70. Calculating the average improvement for all players over both parts shows only a small improvement of 0,79%. The players with extreme outliers showed the highest improvements. However, even though some players improved significantly others could not repeat their good performance from the first part. There were a lot of participants with a drop-in performance in the third part.

Table 47 in the appendix shows the differences for everyone over an average value for the first part (AVG 1) and third part (AVG 3). All values in this table are shown in percentage. AVG´s were calculated based on percentage error for each question and player. On the right side of the table, the difference to 100% is displayed showing how much participants guessed over or under the true value. At the end, differences were calculated with the absolute numbers to find improvements or drops in performance. As mentioned, players that had guessed outside of the boundaries showed a massive improvement. For example, player 3 and 7 (marked in grey). On the other hand, a few players could not use the additional information to their advantage. Seven of those were either close in the first part or just took a huge drop in their performance of over 100% (marked in red). Based on this the overall improvement of all players only sums up to 0,79%. Therefore, no significant prove has been found that additional information on borders for the scale will lead to an improvement of the participants judgments. With a bit more than half of the players.

Table 21 and 22 display the outperformance of the average means in the first part, which resulted then in an average outperformance for all three methods. As table 39 and 40 display the same outperformance for part three. Therefore, the following average outperformances were calculated.

Parts	1	2	3	Difference 1 -> 3
Arithmetic mean	74,81%	68,46%	76,15%	1,35%
Geometric mean	83,85%	57,50%	69,42%	-14,42%
Median	78,85%	58,46%	65,96%	-12,89%
Range	9,04%	10,96%	10,19%	15,77%

Table 20 Outperformance of the averaging methods and improvement

Table 48 now shows all the averages of the methods and furthermore the differences from part one and three. As exhibited, only the arithmetic mean could improve and both other methods took a decrease in their outperformance over the individuals. The geometric mean lost the most performance with -14,41% and generated with the improvement of the arithmetic mean a range of 15,77% for the difference between part one and three. In this table, it is clear to see that close scales have a negative effect on the performance of geometric mean and median a lot.

To sum up the change of the individual judgments in table 47 showed both improvements and drops in performance. The best improvements were shown from players, which had guesses outside of the boundaries and used the additional information to adjust their judgment. Others had close guesses to the true value within the first part and could not repeat the performance within the third part and a few of those had performance losses of over 100%. A similar picture shows table 48 for the averaging methods. Arithmetic mean could improve performance. Geometric mean and median lost performance compared to the first part. Arithmetic mean could improve slightly the average outperformance over the individuals. This shows that wisdom of the crowd can be valuable in situations where individuals have only limited information or knowledge about the task. Even with additional information, wisdom of the crowd generates high numbers of performance of individual judgment. Within part two, averaging methods could only outperform around half of the participants. This showed an impact of percentage scale, which also can be seen in table 48.

Analysis of percentage scale

Looking at the percentage scale on individual level revealed an interesting picture. Question 16 had a true value of only 2% and created high percentage error for almost every participant. It had a max percentage error of 4250% and an average error for all participants of 1192,12%. Therefore, to reduce the impact of only one question it was removed to calculate a new average (AVG – Q16). Results are listed in table 49 from the appendix.

Even though there were no outliers within the data set, it can be said that question 16 showed tenancies of spikes. It was the only question with high percentage errors. The highest error within all other questions was around 368,42%. This resulted in a calculated difference of 823,69% which was high and lead to exclusion of question 16. After recalculating the average percentage error without question 16, differences between the two averages were calculated. First, the difference to 100% was used to look for over- or underestimates from participants. Then absolute difference was measured to see improvements of the individuals. Thirty participants had improved accuracy with a summarized improved result of 4598,38%. Six participants (marked grey) had an enormous upgraded of their judgment with over 238%. Compared to other participants which judgments did not improve with a summarized result of only -354,73%. Therefore, it can be said that without question 16 the individual judgments would have been better off which can also be calculated in a value around 4243,65%.

Results

Collected data showed only a few missing answers mainly from one participant. All other data was useable and was analysed regarding to min and max values of each part. Spikes were found only in part one by comparing the data to possible max and min values of the scale. Based on the ranking system participant 12 could reach the first place overall. The ranking system showed the quality over quantity fact with a few examples within the analysis of each single part. Distribution of points showed a nearly linear increase expect at two points. At those two points, one at the start and one at the end small jumps happened.

Part one showed worst guesses with the highest distance to the true value. As a possible reason was identified. The absence of knowledge over the limits or possible limits of a scale leads to outliers and makes it hard for participants to guess. The analysis of average methods showed the geometric average had the best AE´s in part one. It demonstrates the fact that, geometric mean is not as sensitive to outliers compared to the arithmetic mean. Similar results were found for median, which could slightly outperform the arithmetic mean in part one.

Within part two, the best guesses came close to the true value and some even hit with zero error. Not a single outlier was found. Furthermore, average methods were analysed, and results showed that arithmetic average had the best AE´s overall. A possible reason for this could be that without outliers' geometric mean did not perform as well as in part one. The median had a perfect hit, but its average performance was only slightly better than of the average geometric mean. Performances of all averaging methods in this part were not particularly great. All methods had a drop in their performance, which is displayed in table 48.

Part three showed improvements on both the best and worst guesses. Within worst guesses even without spikes outside of the boundaries from the scale. This happened because additional information was given to all players of the experiment. Arithmetic mean had the best performance of the three averaging methods in part three. Median and geometric mean came close but arithmetic mean had three perfect performances within part three.

Comparing part one to three showed that some participants could improve their guesses while others could repeat their good performance. An important factor was the removal of spikes. By eliminating outliers, performance of geometric mean and median could hold up to the performance of arithmetic mean. In addition, the ranking showed complete different players at the top for both parts. The changes of the individual judgments showed improvements and drops in performance. Highest improvements came from players with guesses outside of the scale compared to a few players, which had drops in their performance of over 100%. Of all three average methods, only the arithmetic mean could improve to some extent. Geometric mean and median drop their performance compared to the first part. Overall

arithmetic mean showed the highest consistent performance and highest overall average performance with 73,14% outperformance to the individuals calculated over all parts.

Analysis of wisdom of the crowd showed its strongest performance with all averaging methods was in part one. The effect of scale can be seen in table 48, which provides all the outperformance numbers over each part. The wisdom of the crowd performance has its lowest point within part two for the percentage scale. Part three differs from part one with the absent of spikes. These spikes did improve performance for geometric mean and median in part one. Only arithmetic mean could hold a constant performance through the whole experiment. Therefore, it was the best averaging method for this experiment overall. Giving additional advice cannot be classified as a positive influence on the individual level, because around half the players improved and the other half decreased their performance in part three. The overall level of judgment only increased around 0,79%. Nevertheless, the biggest positive effect of the additional information was to eliminate outliers completely in part three which resulted into those changes of each averaging method.

The percentage scale showed one surprise with question 16. Participants did not judge nearly with the same accuracy compared to other questions of part two. Therefore, a calculation was made to look at part two without question 16. Results displayed very high improvements compared to only slightly drops in performance for the individual levels. Overall, without question 16 the result could have been improved quite a lot for most individuals.

General Discussion

In the last years, wisdom of the crowd has shown many possibilities for diverse practices and research topics. Because of this, influence factors should to be monitored ensuring processes or experiments can be used to their full potential and reveal unbiased wisdom of the crowd. Furthermore, looking into new possible influence factors should be a part for new research ideas. Ideally, this paper should contribute to this matter. Scales provide an easy way to collect, convert or analyse data from many individuals. Therefore, the influence those scales is important to be recognised and further researched in future studies. In this paper, once again the averaging methods showed a high advantage over most individual

guesses. The greatest effect of averaging methods for wisdom of the crowd can be seen in part one where participants had no knowledge over the exact extent of the kilometre scale. Table 48 summarizes all outperformances and shows this just mentioned fact with the highest numbers for averaging methods in part one. Difference from part one and three are based on the additional advice. This information did not have a pure positive impact on the individual level. As mentioned, the outliers could be eliminated but some other participants had a drop in their performance. Therefore, it is possible that those players got unsecure about their guesses in part one and after receiving information about the range of the scale, their adaptions went into the wrong direction. Additionally, only arithmetic mean could improve its performance, due the lack of outliers both geometric mean and median dropped performance compared to part one. Which leads to the statement that giving out information about the range of an open scale does not only have a positive influence on wisdom of the crowd. Similarly, to that is the influence of an outlier question in the percentage scale. This would decrease wisdom of the crowd because most participants did not even get close to the true value. This could result into a deformation of the whole picture. General wisdom of the crowd was a topic with great-unknown potential and is now a common research topic. The produced results and found influence factors could create a map of now possible research strategies of wisdom of the crowd. The limitation of this work creates opportunities for further research to find more influence factors and contribute to the progress of creating a complete picture of wisdom of the crowd.

Abstract

This paper focused was to draw attention to already know influence factors of wisdom of the crowd and research the influence of scale on wisdom of the crowd. Starting by an overview of already known influence factors of wisdom of the crowd. Followed by a short review of the details within those papers. In my opinion, looking at wisdom of the crowd holds such immense potential to gain knowledge, support, information or cooperation from many individuals and can convert these things into impressive accurate numbers. Experience has shown me, that most people are willing to help but sometimes unable to give the right answers. This behaviour could be very useful with the help of wisdom of the crowd. Because lack of data and biased answers from participants are obstacles for the use of wisdom of the crowd. To ensure a good use of wisdom of the crowd those obstacles must be overcome. Therefore, testing for an unresearched influence factor builds the core element of this work. With the help of participants an experiment of potential influence of the two scale (percentage as closed scale and kilometre as open scale) was done. This should provide another work to the list of the first part and improve understanding for influence factors of using wisdom of the crowd. The performance of the participants was measured with through the best and worst analysis, which translated the number of good placements into a ranking system. Afterwards, analysis for each part of the experiment were done. The arithmetic mean lead to an overall 73,14% outperformance over individual guesses in the experiment showing the power of wisdom of the crowd. Which shows that people can provide information or advise even if their accuracy is off, there input still can be useful by means of wisdom of the crowd this data could lead to an overall more accurate result in the end.

Abstract (Deutsch)

Der Fokus dieser Arbeit lag darauf Aufmerksamkeit auf die Einflussfaktoren von wisdom of the crowd und Forschung über den Einfluss von Skalen auf wisdom of the crowd zu ziehen. Den Beginn machte die Übersicht über bereits bekannte Einflussfaktor auf wisdom of the crowd. Gefolgt von einem kurzen Überblick von den Details aus den Artikeln. Meiner Meinung nach hält wisdom of the crowd immenses Potential in der Gewinnung von Wissen, Unterstützung, Informationen oder auch Kooperation durch viele einzelne Individuen und kann diese Dinge in beeindruckend genaue Zahlen umwandeln. Erfahrung hat mir gezeigt, dass die meisten Menschen gewillt sind zu helfen aber manchmal nicht in der Lage sind die richtige Antwort zu geben. Mit Hilfe von wisdom of the crowd könnte dieses Verhalten sehr nützlich sein. Denn das fehlen von Daten und die Voreingenommenheit bei Antworten sind Hindernisse bei der Nutzung von wisdom of the crowd. Um eine gute Nutzung von wisdom of the crowd zu gewährleisten müssen diese Hindernisse überwunden werden. Deshalb bildet ein noch nicht erforschter Einflussfaktor den Kern dieser Arbeit. Mit der Hilfe von Teilnehmern wurde ein Experiment durchgeführt um den potenziellen Einfluss von zwei Skalen (Prozent als geschlossene Skala und Kilometer als offene Skala) durchgeführt. Dies soll ein zusätzliches Werk für die angeführte Liste im ersten Teil der Arbeit bieten, sowie das Verständnis von Einflussfaktoren über wisdom of the crowd verbessern. Die Performance der Teilnehmer wurde durch die „besten und schlechtesten“ Analyse gemessen, welche die Anzahl der guten Platzierungen in das Ranglisten System übertragen hat. Dadurch wurden die drei Gewinner des Experiments, sowie die beiden niedrigsten Punkte erkannt. Danach wurde die Analyse für die einzelnen Teile gemacht. Bei Benutzung des arithmetischen Mittelwerts konnte eine Gesamtmenge von 73,14% der Teilnehmer übertroffen werden und zeigte damit die Stärke wisdom of the crowd. Dies zeigt, dass Leute Informationen oder Rat zur Verfügung stellen können selbst wenn deren Genauigkeit etwas daneben liegt, können diese Informationen mit Hilfe von wisdom of the crowd im Endeffekt zu einem genaueren Ergebnis führen.

Sources

Arlinghaus, R., & Krause, J. (2013). Wisdom of the crowd and natural resource management. *Trends in Ecology & Evolution, Vol.28*, pp. 8-11.

Chowdhury, W., Burt, C., Akkaoui, A., & Davies, J. (2015). Quanty: An online game for eliciting the wisdom of the crowd. *Computers in Human Behavior, Vol. 49*, pp. 213-219.

Herzog, S., & Hertwig, R. (2013, September 9). Think Twice and then: combining or choosing in dialectical bootstrapping? *Journal of experimental psychology: learning, memory, and cognition. Advance online publication. doi:10.1037/a0034054*.

Herzog, S., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, Vol. 18(10)*, pp. 504-506.

Juni, M. Z., Eckstein, M. P., & Enns, J. T. (2015). Flexible Human Collective Wisdom. *Journal of Experimental Psychology, Vol. 41(6)*, pp. 1588-1611.

Larrick, R. P., & Soll, J. B. (2006). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle. *Management Science, 2006, Vol.52(1)*, pp. 111-127.

Lee, M., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition, Vol39(5)*, pp. 914-923.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences Vol. 108(22)*, pp. 9020-9025.

Mannes, A. E. (2009). Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science, Vol.55(8)*, pp. 1267-1279.

Mannes, A. E., Soll, J. B., Larrick, R. P., Simpson, J. A., & Kawakami, K. (. (2014). The Wisdom of Select Crowds. *Journal of Personality and Social Psychology, Vol. 107(2)*, pp. 279-299.

Nickerson, J. V., Zahner, D., Corter, J. E., Tversky, B., & Yu, L. (2009). Matching Mechanisms to Situations Through the Wisdom of the Crowd. *Thirtieth International Conference on Information Systems*, (pp. 1-15). Phoenix.

Prelec, D., Seung, H. S., & Mccoy, J. (2017). A solution to the single-question crowd wisdom. *Nature, Vol. 541(7638)*, pp. 532-535.

Riedl, C., Blohm, I., Leimeister, J. M., & Krcmar, H. (2010). Rating scales for collective intelligence in innvoation communities: Why quick and easy decision making does not get it right. *Thirty First International Conference on Information Systems*, (pp. 1-21). St. Louis.

Shankar Mysore, A. Y. (2015). Investigating the "Wisdom of Crowds" at Scale. *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, (pp. 75-76).

Soll, J. &. (2009). Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition 35(3)*, pp. 780-805.

van Dolder, D., & van den Assem, M. (2018). The wisdom of the inner crowd in three large natural experiments. *Nat Hum Behav, Vol. 2(1)*, pp. 21-26.

Online Sources

https://www.google.com/search?client=firefox-b-d&q=lenght+of+the+amazonas

https://en.wikipedia.org/wiki/Equator (visit time 22.01.2020, 10:37)

Appendix

Q-Nr.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Max	32.000	24.000	28.000	30.000	40.000	55.000	50.000	32.000	45.000	30.000
Min	900	300	200	200	780	630	100	800	1.000	400
Q-Nr.	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30
Max	12.000	19.000	10.000	11.000	18.000	8.000	18.000	11.000	19.861	9.000
Min	600	200	300	80	150	500	1.000	500	1.300	100

Table 21 Min and max for questions with the open scale (km)

Placement	Nr.	Q1	True Value	AE
First	25	6.000	6.249,28	249,28
	30	6.000	6.249,28	249,28
	39	6.000	6.249,28	249,28
	40	6.000	6.249,28	249,28
	52	6.000	6.249,28	249,28
Second	9	6.500	6.249,28	250,72
Third	46	5.500	6.249,28	749,28

Table 22 Placements of participants for question 1

Player	Last	Second to Last	Total placements at the end
7	9	4	13
3	6	6	12
35	3	1	4
42	1	3	4
45	2	2	4

Table 23 Number of placements at the end

Player	First	Second	Third	Total number of top 3	Total Points
3	2	2	3	7	15
35	1	2	1	4	9
7	2	0	0	2	8
42	1	0	3	4	7
45	0	1	2	3	4

Table 24 Participants with highest number of placements at the end

Player	First	Second	Third	Total number of top 3	Total Points
11	0	1	1	2	3
50	0	1	1	2	3
19	0	0	2	2	2
33	0	1	0	1	2

Table 25 Worst participants

Part one of the experiment

	Q1	Q2	Q3	Q4	Q5
Best	6.000,00	2.500,00	3.000,00	3.207,00	8.000,00
True Value	6.249,28	2.558,89	3.128,34	3.361,78	8.277,15
Best AE	249,28	58,89	128,34	154,78	277,15

Table 26 Best guesses for questions of first part (1/2)

	Q6	Q7	Q8	Q9	Q10
Best	15.000,00	7.500,00	6.700,00	6.800,00	2.101,00
True Value	14.243,75	7.276,66	6.670,58	6.400,68	2.094,44
Best AE	756,25	223,34	29,42	399,32	6,56

Table 27 Best guesses for questions of first part (2/2)

	Q1	Q2	Q3	Q4	Q5
Worst	32.000,00	24.000,00	28.000,00	30.000,00	40.000,00
True Value	6.249,28	2.558,89	3.128,34	3.361,78	8.277,15
Worst AE	25.750,72	21.441,11	24.871,66	26.638,22	31.722,85

Table 28 Worst guesses for questions of first part (1/2)

	Q6	Q7	Q8	Q9	Q10
Worst	55.000,00	50.000,00	32.000,00	45.000,00	30.000,00
True Value	14.243,75	7.276,66	6.670,58	6.400,68	2.094,44
Worst AE	40.756,25	42.723,34	25.329,42	38.599,32	27.905,56

Table 29 Worst guesses for questions of first part (2/2)

	Q1	Q2	Q3	Q4	Q5
Arithmetic AE	168,78	836,71	28,83	1175,85	1247,85
Geometric AE	1.093,87	305,30	1.089,02	825,84	291,98
Median AE	499,28	158,89	628,34	511,78	622,85

Table 30 Comparison of averaging methods AE's for questions of first part (1/2)

	Q6	Q7	Q8	Q9	Q10
Arithmetic AE	4.523,54	2.673,70	760,02	1765,70	1682,50
Geometric AE	7.034,58	4.839,02	728,49	40,23	31,42
Median AE	6.743,75	4.075,66	670,58	699,32	94,44

Table 31 Comparison of averaging methods AE's for questions of first part (2/2)

Part two of the experiment

	Q11	Q12	Q13	Q14	Q15
Best	30,00	71,00	65,00	40,00	40,00
True value	30,00	71,00	65,00	40,00	40,00
Best AE	0,00	0,00	0,00	0,00	0,00

Table 32 Best guesses for questions of second part (1/2)

	Q16	Q17	Q18	Q19	Q20
Best	2,00	19,00	80,00	45,00	80,00
True value	2,00	19,00	82,00	44,00	82,00
Best AE	0,00	0,00	2,00	1,00	2,00

Table 33 Best guesses for questions of second part (2/2)

	Q11	Q12	Q13	Q14	Q15
Worst	90,00	45,00	35,00	80,00	81,00
True value	30,00	71,00	65,00	40,00	40,00
Worst AE	60,00	26,00	30,00	40,00	41,00

Table 34 Worst guesses for questions of second part (1/2)

	Q16	Q17	Q18	Q19	Q20
Worst	85,00	70,00	25,00	96,00	16,00
True value	2,00	19,00	82,00	44,00	82,00
Worst AE	83,00	51,00	57,00	52,00	66,00

Table 35 Worst guesses for questions of second part (2/2)

	Q11	Q12	Q13	Q14	Q15
Arithmetic AE	13,00	0,69	0,83	11,52	3,31
Geometric AE	17,87	1,09	0,56	19,40	10,45
Median AE	18,00	1,00	0,00	10,00	10,00

Table 36 Comparison of averaging methods AE´s for questions of second part (1/2)

	Q16	Q17	Q18	Q19	Q20
Arithmetic AE	21,84	6,49	16,26	8,67	27,12
Geometric AE	12,53	13,16	18,43	4,26	31,92
Median AE	13,00	12,25	12,00	7,00	22,00

Table 37 Comparison of averaging methods AE´s for questions of second part (2/2)

Part three of the experiment

	Q21	Q22	Q23	Q24	Q25
Best	3.000,00	8.000,00	3.500,00	4.000,00	7.000,00
True Value	2.948,97	8.108,88	3.599,07	3.971,85	7.150,00
Best AE	51,03	108,88	99,07	28,15	150,00

Table 38 Best guesses for questions of third part (1/2)

	Q26	Q27	Q28	Q29	Q30
Best	1.200,00	6.800,00	1.000,00	6.500,00	2.900,00
True Value	1.142,16	6.401,22	1.004,59	6.732,12	2.804,94
Best AE	57,84	398,78	4,59	232,12	95,06

Table 39 Best guesses for questions of third part (2/2)

	Q21	Q22	Q23	Q24	Q25
Worst	12.000,00	19.000,00	10.000,00	11.000,00	18.000,00
True Value	2.948,97	8.108,88	3.599,07	3.971,85	7.150,00
Worst AE	9.051,03	10.891,12	6.400,93	7.028,15	10.850,00

Table 40 Worst guesses for questions of third part (1/2)

	Q26	Q27	Q28	Q29	Q30
Worst	8.000,00	18.000,00	11.000,00	19.861,00	9.000,00
True Value	1.142,16	6.401,22	1.004,59	6.732,12	2.804,94
Worst AE	6.857,84	11.598,78	9.995,41	13.128,88	6.195,06

Table 41 Worst guesses for questions of third part (2/2)

	Q21	Q22	Q23	Q24	Q25
Arithmetic AE	477,01	566,17	10,70	1242,81	992,06
Geometric AE	247,27	2107,77	982,58	2354,55	233,15
Median AE	51,03	1108,88	599,07	1971,85	850,00

Table 42 Comparison of averaging methods AE´s for questions of third part (1/2)

	Q26	Q27	Q28	Q29	Q30
Arithmetic AE	1091,23	240,40	2637,63	2710,06	47,69
Geometric AE	611,56	1387,66	1596,79	1799,44	777,61
Median AE	557,84	1401,22	1995,41	2267,88	804,94

Table 43 Comparison of averaging methods AE´s for questions of third part (2/2)

Comparing part one and three

Number of differences below zero	30,00	
Number of differences above zero	21,00	
Number without a difference	1	
		Averages
Negative difference	- 102.700,00	- 3423,33
Positive difference	69.910,33	3329,06
Overall difference	- 32.789,67	- 94,27

Table 44 Differences of part one and three for the individual level

Question	True Value
Q5	8.277,15
Q8	6.670,58
Q9	6.400,68

Question	True Value
Q22	8.108,88
Q29	6.732,12
Q27	6.401,22

Table 45 Comparison of questions with similar true values

	Arithmetic AE	Geometric AE	Median AE		Arithmetic AE	Geometric AE	Median AE
Q5	1.247,85	291,98	622,85	Q22	566,17	2.107,77	1.108,88
Q8	760,02	728,49	670,58	Q29	2.710,06	1.799,44	2.267,88
Q9	1.765,7	40,23	699,32	Q27	240,4	1.387,66	1.401,22

Table 46 Comparison of averaging methods AE´s part one to three

	Arithmetic AE	Geometric AE	Median AE		Arithmetic AE	Geometric AE	Median AE
Q5	80,77%	96,15%	94,23%	Q22	94,23%	61,54%	78,85%
Q8	86,54%	86,54%	86,54%	Q29	57,69%	65,38%	57,69%
Q9	57,69%	100,00%	80,77%	Q27	100,00%	86,54%	73,08%

Table 47 Comparison of averaging methods percentage outperformance

Analysis of individuals

Average percentage error of individuals		
Player	AVG 1	AVG 3
1	92,95	103,12
2	169,50	103,05
3	486,01	239,65
4	161,31	134,58
5	153,12	181,87
6	165,79	196,72
7	639,57	233,87
8	102,47	170,39
9	83,59	85,26
10	104,65	232,82
11	83,35	92,34
12	95,84	100,61

Difference to 100%		Difference
AVG 1	AVG 3	
-7,05	3,12	3,93
69,50	3,05	66,45
386,01	139,65	246,36
61,31	34,58	26,73
53,12	81,87	-28,75
65,79	96,72	-30,93
539,57	133,87	405,70
2,47	70,39	-67,92
-16,41	-14,74	1,67
4,65	132,82	-128,17
-16,65	-7,66	8,98
-4,16	0,61	3,55

13	136,16	236,28	36,16	136,28	-100,12
14	124,65	137,00	24,65	37,00	-12,35
15	63,39	63,26	-36,61	-36,74	-0,12
16	133,70	93,86	33,70	-6,14	27,57
17	103,22	212,72	3,22	112,72	-109,49
18	114,05	219,54	14,05	119,54	-105,49
19	37,42	184,13	-62,58	84,13	-21,55
20	49,03	110,55	-50,97	10,55	40,42
21	125,45	238,54	25,45	138,54	-113,08
22	133,69	191,41	33,69	91,41	-57,72
23	134,17	114,71	34,17	14,71	19,47
24	75,86	140,67	-24,14	40,67	-16,53
25	115,97	96,59	15,97	-3,41	12,56
26	59,78	76,43	-40,22	-23,57	16,65
27	107,62	131,78	7,62	31,78	-24,16
28	169,74	153,63	69,74	53,63	16,11
29	82,61	143,46	-17,39	43,46	-26,07
30	69,74	111,44	-30,26	11,44	18,82
31	50,20	99,61	-49,80	-0,39	49,41
32	98,68	204,32	-1,32	104,32	-103,00
33	35,18	58,93	-64,82	-41,07	23,75
34	87,69	158,17	-12,31	58,17	-45,86
35	49,30	107,98	-50,70	7,98	42,72
36	73,28	82,99	-26,72	-17,01	9,71
37	108,25	109,26	8,25	9,26	-1,01
38	23,72	34,93	-76,28	-65,07	11,22
39	99,03	57,07	-0,97	-42,93	-41,95
40	70,12	74,66	-29,88	-25,34	4,54
41	159,24	69,56	59,24	-30,44	28,80
42	11,12	264,99	-88,88	164,99	-76,11
43	53,88	100,08	-46,12	0,08	46,04
44	53,76	113,28	-46,24	13,28	32,96
45	187,89	313,75	87,89	213,75	-125,86
46	97,56	76,58	-2,44	-23,42	-20,98
47	107,21	160,69	7,21	60,69	-53,48
48	53,76	111,12	-46,24	11,12	35,12
49	33,07	69,67	-66,93	-30,33	36,60
50	38,57	94,27	-61,43	-5,73	55,71
51	118,00	99,16	18,00	-0,84	17,16
52	123,13	120,30	23,13	20,30	2,83

Table 48 Individual differences from part one to three

Analysis of percentage scale

Average percentage error of individuals		
Player	AVG	AVG - Q16
1	313,30	70,33
2	296,37	107,08
3	208,53	65,04
4	249,51	110,56
5	109,81	66,45
6	78,09	70,10
7	261,90	96,55
8	267,07	74,52
9	552,62	141,80
10	116,39	73,76
11	87,44	69,37
12	98,53	87,25
13	91,73	74,15
14	144,34	77,04
15	213,65	70,72
16	63,45	59,39
17	102,23	91,36
18	109,83	66,48
19	115,66	100,73
20	415,01	100,01
21	90,88	73,20
22	80,88	50,98
23	449,23	110,26
24	339,46	99,39
25	78,42	59,35
26	63,24	53,60
27	282,09	91,21
28	125,32	83,69
29	372,11	80,13
30	102,06	57,85
31	347,69	108,55
32	221,55	118,39
33	131,34	62,60
34	277,20	85,78
35	155,23	116,92
36	227,20	85,78
37	89,49	82,77
38	198,96	82,18
39	101,67	85,19
40	98,14	64,60
41	235,67	95,18

Difference to 100%		Difference
AVG	AVG - Q16	
213,30	-29,67	183,63
196,37	7,08	189,29
108,53	-34,96	73,57
149,51	10,56	138,94
9,81	-33,55	-23,74
-21,91	-29,90	-7,99
161,90	-3,45	158,45
167,07	-25,48	141,60
452,62	41,80	410,82
16,39	-26,24	-9,85
-12,56	-30,63	-18,06
-1,47	-12,75	-11,27
-8,27	-25,85	-17,59
44,34	-22,96	21,38
113,65	-29,28	84,37
-36,55	-40,61	-4,06
2,23	-8,64	-6,41
9,83	-33,52	-23,68
15,66	0,73	14,93
315,01	0,01	315,00
-9,12	-26,80	-17,68
-19,12	-49,02	-29,90
349,23	10,26	338,97
239,46	-0,61	238,85
-21,58	-40,65	-19,06
-36,76	-46,40	-9,64
182,09	-8,79	173,29
25,32	-16,31	9,01
272,11	-19,87	252,24
2,06	-42,15	-40,09
247,69	8,55	239,15
121,55	18,39	103,16
31,34	-37,40	-6,06
177,20	-14,22	162,98
55,23	16,92	38,31
127,20	-14,22	112,98
-10,51	-17,23	-6,72
98,96	-17,82	81,14
1,67	-14,81	-13,13
-1,86	-35,40	-33,54
135,67	-4,82	130,85

42	373,91	121,01	273,91	21,01	252,90
43	78,80	70,89	-21,20	-29,11	-7,91
44	80,94	62,16	-19,06	-37,84	-18,78
45	150,09	83,43	50,09	-16,57	33,51
46	112,30	69,23	12,30	-30,77	-18,47
47	233,58	65,08	133,58	-34,92	98,66
48	414,70	99,66	314,70	-0,34	314,36
49	115,80	73,11	15,80	-26,89	-11,09
50	306,01	90,01	206,01	-9,99	196,03
51	91,44	91,60	-8,56	-8,40	0,16
52	191,36	101,52	91,36	1,52	89,85

Table 49 Percentage scale differences without question 16