Erasmus University Rotterdam

Erasmus School of Economics
Master Specialization Behavioral Economics

MASTER THESIS

Wisdom of the Crowd: Comparison of the CWM, Simple Average and Surprisingly Popular Answer Method

Author: **Bc. Kristyna Matoulkova**

Supervisor: **M. Phil. Benjamin Tereick**

Academic Year: **2017/2018**

Declaration of Authorship

The author hereby declares that she compiled this thesis independently, using only the listed resources and literature.

The author grants to Erasmus University Rotterdam permission to reproduce and to distribute copies of this thesis document in whole or in part.

Rotterdam, September 18, 2017

Signature

Acknowledgments

I would like to sincerely thank my supervisor M. Phil. Benjamin Tereick for his support, useful recommendations and advice that he gave me thorough the writing process.

Abstract

On a daily basis people and institutions are forced to make optimal choices, provide correct solutions to many times very complex tasks and predict future outcomes. Since individual answers tend to suffer from random errors in judgment, it is frequently better to rely on mathematical combination of people's opinions in order to make the best decision. This approach, which is academically called the wisdom of the crowd, lowers the random error and delivers an accurate answer that is most of the times better than any individual guess. In this thesis I compare the performance of three aggregation methods withing the crowd wisdom concept: a simple average (SA), a surprisingly popular answer algorithm (SPA) and a contribution weighted model (CWM). Besides that I also conduct a supplementary research including the BTS score performance comparison, regression of the CWM score on the BTS score and an analysis of cheating behavior. By means of paper and online quiz about general knowledge questions I collect data, compare the outcomes of the models and conduct several robustness checks. The main findings of the paper show that both the SPA and the CWM perform significantly better than the SA for large samples, however, the results are more robust for the SPA/SA case. Moreover, there is no difference in the performance of the SPA and the CWM. Therefore, I conclude that the SPA is the most appropriate for general use as it has less requirements regarding the type and the number of the questions than the CWM and its good performance is more robust. Finally, further research should be focused on better incentivization of subjects so that more questions can be used for the analysis.

JEL Classification	D70, D90
Keywords	crowd wisdom, contribution weighted model, surprisingly popular answer, simple average, performance comparison

Author's e-mail	`k.matoulkova@seznam.cz`
Supervisor's e-mail	`tereick@ese.eur.nl`

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Deciding quickly and wisely is essential in order to become successful in the fast-paced environment of the 21^{th} century. Since individual decisions tend to be affected by a random error and a tendency to over or underestimate probabilities (Yaniv, 2004), it is many times better to rely on mathematical combination of multiple opinions when facing problems. With this approach one can reduce the random error in individual judgment and obtain the same or, most of the time, better answer than any individual guess.

In 1906, a British scientist Francis Galton discovered during one of his strolls that people's average guess about a weight of a dead animal was very close to the true weight of the corpse. This finding was striking as before collective decisions had been considered as inefficient and mostly wrong (Surowiecki 2005). By his discovery, Galton gave rise to a new academic focus analyzing the wisdom of the crowd that has become one of the core parts of behavioral science. Many studies had followed including a famous bean jar experiment from Jack Treynor (1987). Nowadays, it is mainly institutions who use the crowd wisdom, namely through prediction markets, prediction polls and crowd-sourcing to forecast demand and increase profits. Besides the simple average that Galton used, new aggregation methods and algorithms have emerged that make the crowd wisdom a very powerful tool for decision-making.

Two of the crowd wisdom methods are especially worth mentioning: a surprisingly popular answer method (SPA) developed by Prelec et al. (2017) and a contribution weighted model (CWM) derived by Budescu and Chen (2014). The SPA determines the best answer as the one that is more frequent than expected. It excels especially in cases when being familiar with an item does not lead to the right solution. In other words, it delivers a correct answer in

a question such as: Is the capital of Australia Canberra or Sydney?, where the simple average answer would most probably fail. On the other hand, the CWM uses the individual relative contribution to the group performance to determine experts in the crowd and employs the combination of their answers to define the best solution. Thus, the people involved in the model are those who are mostly right when the rest of the group is wrong. In past academic research, the SPA has outperformed many different methods including the confidence weighting algorithm, however, it has not been directly compared to the CWM that is also based on the subjective confidence and performs well with large sample sizes and various question topics. As there is a gap in academic research on this matter, this paper compares the performance of the CWM and SPA, namely in the domain of general knowledge. The simple average method is also included in the analysis. As it is expected to perform worse than the two models, it serves as a sort of check that both the SPA and the CWM are properly constructed. Additionally, the paper also covers a supplementary analysis focused on the relationship between the Bayesian Truth Serum score Prelec (2004) and the CWM score, cheating behavior and several robustness checks. Data for this thesis were collected via paper and online quiz.

This thesis has the following structure: Chapter 2 provides a literature review on the topic of the crowd wisdom, namely, it focuses on the development and applications of the concept, 4 conditions to make the group wise and on aggregation methods and research hypotheses. Subsequently, Chapter 3 presents the methodology used in the paper: it covers the explanation of the CWM and SPA, experimental design, data description as well as data analysis. Moreover, it explains different robustness checks and supplementary analyses used in the paper. Chapter 4 is focused on results and discussion. Finally, Chapter 5 concludes the paper by summarizing the most important points.

Chapter 2

Literature Review

This chapter will first briefly introduce the development and modern applications of the crowd wisdom. Subsequently, it will discuss four main conditions under which the concept works based on Surowiecki (2005). Finally, the most important aggregation methods will be outlined and a research question presented.

2.1 Development and Applications

Nowadays, in the period of ubiquitous stress and time constraints, people and institutions are frequently forced to solve problems quickly and efficiently. In such situations, they can rely on the answer of a single person which may, however, suffer from random error in judgment and various biases. This was proven, for instance, by Jain et al. (2015) who showed that biases such as anchoring and disposition effect negatively influence decisions of individual investors, namely their timing of a contract and purchasing shares. Additionally, they noted that even if debiasing techniques are available, they tend to be time-consuming, take a lot of effort and their application is not always efficient.

Therefore, if a single person is not always the best source of information, one can try to collect more opinions, mathematically combine them in order to mitigate individual noise and obtain a better answer. In 1906, Francis Galton, a famous French scientist, happened to be a witness of a guessing competition in a local market in a small French city. Butchers, farmers but also random passers-by were writing on a piece of paper their best estimations about the weight of a slaughtered ox. When Galton collected these guesses after the competition, he noticed that the simple average of all the estimations was only 1 pound away

from the true weight of the ox. This surprising finding subsequently gave rise to the study of the crowd wisdom (Surowiecki 2005).

One of the first laboratory experiments directly testing this concept was conducted by Hazel Knight in the third decade of the 20th century. The researcher asked his students to predict the temperature in his classroom. Their aggregated group guess was about 80 percent better than individual estimates, which clearly supported the crowd wisdom approach (Sunstein and Hastie 2015). Another early experiment was conducted by Kate Gordon, also in 1920s. She took ten same-size glasses filled with different amounts of cotton and shots and asked 200 students to rank these glasses based on their weight. As the difference between the heaviest and the lightest tumbler accounted only for 1.6 gram, individual estimates were substantially inaccurate. On the other hand, an aggregated guess of the whole group yielded 91% correct order (Gordon 1924). Probably the most well-known experiment on the group knowledge concerned a bean jar contest and was conducted in 1980s. A finance professor, Jack Treynor, asked his students to estimate the number of beans in two different jars. The average guess was only 31 and 21 beans away from the true values of 810 and 850 beans, respectively. All together only three individual predictions were better than the average group guess (Treynor 1987).

The wisdom of the crowd moved from successful laboratory experiments to a real world settings in winter 1986 after the space shuttle Challenger exploded. At that time, two scientists Maloney and Mulherin noticed that right after the accident, the stock price of one of the four main space project contractors fell substantially more than the prices of the other three. In other words, the stock market participants collectively predicted by their demand that this one 'loosing' company was guilty from the collision, which eventually turned out to be true (Surowiecki 2005). Two years later, the College of Business in Iowa came up with an idea of using group knowledge to predict the outcome of American elections. Therefore, they designed the Iowa Electronic Market (hereafter IEM), the first type of prediction market - also called information market (Forsythe et al. 1992). In these markets, participants trade with contracts whose monetary outcomes depend on events happening in the future. For instance, a contract would pay $1 to its owner if football team A wins a championship next year and $0 otherwise. Based on the assumption of market efficiency, the price of this contract will deliver the best crowd's prediction about what the real outcome of the given event will be. Thus, a deal which is traded for 80 cents on the dollar means that traders predict the prob-

ability of A team's triumph is 80 percent (Wolfers and Zitzewitz 2004). The Iowa College used the IEM for the first time to predict the margin of George Bush's victory in 1988 presidential elections. The trial resulted in a great success as the forecasted values were only one percentage point away from the real results (Forsythe et al. 1992). Since then, this market has been frequently applied and is still active for research purposes nowadays. Furthermore, the success of the IEM gave rise to other prediction markets that are currently used especially by large corporations because they help to enhance their overall performance. Namely, they lead to the increased flow of information (lower information costs), encouragement of honest responding by firm monitors as well as to incentivization of employees to act in the best interest of their company (lower agency costs) (Abaramowicz and Henderson 2006). One concrete example could be TagTrade, a prediction market used by American retailer Best Buy for forecasting customer related events. Predictions from this market result in 95 percent high accuracy and thus lead to better business strategies (Gallaugher 2012). Despite the undeniable benefits, information markets have also certain drawbacks. For instance, most of the US state laws restrict gambling on which the concept is based. As a result, the possibilities to take an advantage of this crowd wisdom application are sometimes limited in certain regions (Arrow et al. 2008). On top of that, the market prices can be distorted by several biases that lower their forecasting accuracy. These biases depend, for example, on the trader's level of risk aversion: the lesser it is, the more are prices biased toward the middle, 50%, value (Wolfers and Zitzewitz 2006).

Besides prediction markets, the crowd wisdom has been recently used also in prediction polls. In these polls, individuals or teams provide forecasts about outcomes of unknown future events and compete with each other about who will have a higher accuracy of prediction. The final score that each participant gets is based, for instance, on a Brier scoring rule which gives bonus points for confidence in correct answers and a strict penalty for overconfidence in wrong answers (Tetlock et al. 2017). Atanasov et al. (2016) focused on comparing prediction polls with a performance of prediction markets. In their study, they made a two-year-long tournament in accuracy of predicting geo-political events. People participating in information markets were ranked based on their earnings from double auction while subjects in prediction polls were evaluated according to the Brier score. The authors concluded that prices in the markets were more accurate in forecasts than the simple average of judgments from prediction polls. However, the latter outperformed the former when weighting

based on performance or recalibration was used in the polls as well as when the questions were longer and new to a respondent. Thus, this finding indicated that the polls are under certain conditions more useful than the prediction markets.

Probably the newest sphere where the crowd wisdom concept has been utilized in the 21st century is social media. With constantly increasing popularity of the internet, sites such as Twitter, YouTube, Facebook or Flickr provide a valuable source of data. All the 'Likes' on Facebook, downloaded videos and pictures express some opinions on events. Naturally, gathering this information leads to extensive applications in many spheres. For instance, Jin et al. (2010) used data from an image hosting website Flickr for predictions in marketing, politics and economics. They managed to accurately predict the majority of outcomes including the winner of the US presidential elections in 2008 or the product sales of popular items including iPod. Consequently, Eickhoff and Muntermann (2016) tested the validity of social media knowledge by comparing forecasting accuracy between social media users and professional analysts. They found out that experts, in general, adapted more quickly to new facts, updated their beliefs and thus scored higher in accuracy of predictions. However, if social media users satisfied one or more of the three basic crowd wisdom conditions: independence, diversity and decentralization, their guesses outperformed the experts.

Finally, the concept of crowdsourcing is worth mentioning briefly as it has been quite frequently cited in the connection to the modern crowd wisdom applications. This term is defined as the act of collecting opinions, labor and ideas from many non-experts in order to use them for business or policy purposes, for instance, via professional platforms such as Amazon Mechanical Turk (Parent and Eskenazi 2011). Its main benefits lie in the low cost of participation as well as in low time requirements, which is an advantage especially in comparison with the prediction polls. A typical example of the crowdsourcing platform is online encyclopedia Wikipedia. On this collaborative site, a large number of individuals provide small informative contributions which subsequently aggregate in high-quality information. In other words, the website benefits from the crowd wisdom concept (Lehdonvirta and Bright 2015).

2.2 The Wisdom of the Crowd Principles

After mentioning the development and modern applications of the crowd wisdom, it is natural to think about the conditions under which the concept generally works. Therefore, in this section I will briefly discuss three specific principles leading to a wise group that were defined by Surowiecki (2005): 1) diversity of the crowd which is based, for instance, on gender, age, location and expertise (Hosseini et al. 2015), 2) independence of individuals in the crowd (the opinion of a person in a group should depend only on his or her own beliefs, not on the beliefs of other people), 3) decentralization, meaning that people divide work and subsequently draw on local knowledge.

The first principle, diversity, means in a basic sense that not all the members of the group need to be highly educated, let alone have high IQ, to deliver an optimal aggregated outcome. Actually, it is the other way around: the more different people are, the more creative ideas and solutions they come up with. By taking the average of their diverse - though sometimes extreme opinions - the outliers cancel out and we get the optimal answer (Surowiecki 2005). In an effort to explain this phenomenon more theoretically, Yaniv (2004) noted that an individual guess consists of three different parts: the 'truth', the random error in individual judgment, and the tendency to over or underestimate the probability of an event (systematic error). Based on the statistical analysis, the average of multiple opinions has lower random error than individual opinion, given that under/over-estimating bias is close to zero. Thus a diverse crowd should outperform an individual estimate. Naturally, the importance of the crowd variety has been frequently stressed in the literature. For instance, Arazy et al. (2006) pointed out that the group diversity (along with the group size) is behind the success of Wikipedia because various opinions from different people reduce biases emerging from collective decisions such as conformity and group thinking. Importantly, even though diversity is crucial, it is sometimes costly and time consuming to reach. In order to tackle this issue, Herzog and Hertwig (2009) came up with an idea that instead of averaging estimates of the whole diverse crowd, one can use only one individual and reduce his or her estimation error by dialectical bootstrapping. This method is based on taking the average of first individual estimate and dialectical second estimate that depends on different knowledge than the first one. The results of the study showed that bootstrapping was more accurate than using just the first estimate thus it could be employed in cases when there is a lack of people or too tight budget to

create the diverse crowd. Similarly, Teevan and Yu (2017) asked an individual to look at the task from different perspectives in order to replace the need for the group diversity. Not surprisingly, this role-playing approach resulted in generating more creative ideas in comparison with individuals who did not take part in multiple-thinking. In other words, the two above mentioned papers basically indicated that in the future the crowd wisdom might be replaced by the wisdom of a single individual who would be able to identify himself or herself with different scenarios and come up with diverse perspectives.

The importance of the second principle, independence, is quite straightforward: people are more likely to identify a good decision if they are not influenced by judgments of other members in the group (Hosseini et al. 2015). In the real world, however, reaching an independent judgment might be difficult as almost everybody has access to public information and therefore, the opinions are usually correlated (King et al. 2012). A clear example of what happens when the independence principle is not met was provided by Lorenz et al. (2011) who studied the effect of social influence on decisions of individuals in a group. They conducted an experiment in which participants performed 6 different estimation tasks about crime and geography. Each member had to provide 5 consecutive estimates on each task. In a control group, subjects did not see the answers of the other people in these estimations, in a treatment group, they could either observe the group's average estimate from a previous round or full information of all participant's guesses. The results showed that the treatment group significantly fell behind the control group in the estimations. Thus, based on this finding, social influence distorts the crowd wisdom knowledge. However, King et al. (2012) argued that knowing other people's answers is not always harmful. He showed that if individuals are provided not with the information on the past average estimations but with the current group's best guess, the judgments become closer towards the true value and the crowd performs better. In other words, sequential decisions and subsequent aggregation can suppress the negative effect of social dependence in the crowd.

The third important principle concerning the quality of the crowd is decentralization. If people work on a problem in a decentralized manner, there is a higher chance that they will come up with a good solution than working in a centralized way (with only one person or center being in charge). The main benefit of decentralization is that while it supports independence and deep focus on the task, it also allows for close coordination of activities among people. In an ideal state, different groups of individuals should specialize on different

parts of a problem and gain local knowledge. Afterwards, this knowledge would be aggregated into the collective wisdom of the crowd (Surowiecki 2005).

Importantly, Surowiecki (2005) has also defined a fourth principle, aggregation, which ensures that there exists a way to transfer private opinions into a collective judgment. Since this last principle is tightly connected to the topic of this thesis, I devote it the whole next section of this paper.

2.3 Aggregation Methods

One of the easiest ways of aggregating individuals' judgment is taking a simple average. However, this method has two main drawbacks. Firstly, some people decide irrationally because their opinions are influenced by several biases such as propensity to conform to the opinions' of the others (e.g. Das et al. 2013), or following a wrong intuition when making predictions (e.g. Simmons et al. 2011). In such instances, the group's average knowledge tends to be distorted. Secondly, the expertise of judges can substantially vary in the crowd. In that case, a decision-maker might prefer focusing on guesses of only few, 'better' subjects rather than taking the average in order to get a correct answer. Logically, the two mentioned reasons lead to an effort to improve the crowd wisdom by various weighting methods as well as by choosing a smaller crowd within a larger crowd.

Assigning a particular weight to an individual in the crowd means that we assume he or she has a certain level of expertise relative to the rest of the group. There are several standards according to which one can decide: subjective criteria which is based on how a subject evaluates herself or how peers evaluate the subject, objective criteria (not that frequent) that depends on education or professional experience, and criteria following from empirical evidence, e.g. the performance of a subject in previous experimental tasks (Budescu and Chen 2014). Koriat (2012) analyzed the efficiency of weights based on subjective criteria, namely self-confidence. In four experimental studies, he asked his subjects about general knowledge questions and eye illusions. Some of the tasks were easy because answer options were chosen to represent their domains e.g: The capital of France is Paris/Stockholm. However, there were also tricky questions, where familiarity with the item did not reflect the truth, e.g.: The capital city of Australia is Sydney/Canberra. In the first case, subjects with higher confidence were more frequently correct so that assigning weights based on the level of subjective certainty made sense. Yet, in the second case, higher

confidence did not lead to a true answer and the results were inferior in comparison with the accuracy of individual estimates and the simple average. To conclude, the confidence weights are not always suitable, therefore, they need to be used with caution. Prelec et al. (2017) came up with a solution to the tricky questions mentioned above and defined the best answer as the one that is surprisingly popular. Their algorithm originating from Prelec (2004) worked as follows: firstly, people were provided with a statement about e.g. geography such as: "Chicago is the capital of Illinois", and should decide whether they agree or disagree with it. Secondly, they were asked about what percentage of other people, they thought, endorsed the statement. This method then chose the answer which was in reality more frequent than predicted by the respondents. Interestingly, the procedure worked well for both easy and misleading questions and it outperformed the unweighted model as well as the subjective confidence weighted approach. Aspinall (2010) weighted experts by using the Cooke's method which defines professionals based on past-performance criteria. In his study, the author first used respondents' answers to eleven questions about time-to-failure of specific dams to calculate weights for a new set of 'dam' questions. This approach delivered by far better accuracy estimate than when the simple average was used. Note that even though this 'performance' method has gained large popularity in the past decades, it is certainly not perfect. Subjects' predictions are usually highly correlated in a sense that there will be many cases when everybody performs well in a group as well as cases when all the scores are poor. In such situations, measuring absolute output, such as Aspinall (2010) did, is not discriminatory (no clear differentiation among judges), and using weights based on participant's performance relative to the crowd would be a better option (Budescu and Chen 2014). A completely new weighting approach has been developed by Du et al. (2017) who realized that most of the aggregation models assume independence which, however, does not hold in many cases. Therefore, the authors proposed so called CrowdIQ model that accounted for possible dependence and outperformed previously mentioned Cooke's method as well as the contribution weighted model from Budescu and Chen (2014). The weights were, in this case, based not only on the accuracy of previous guesses but also on when these guesses were made so that judges influenced by the answers of others got lower weights. On the other hand, Wang et al. (2011) focused mostly on solving a problem of respondents' incoherent answering and came up with a weighted coherent adjustment. In their setting, two objective penalties were proposed: incoherence penalty and

consensus deviation in order to estimate individual's credibility in the crowd. The weighted aggregation model based on these measures outperformed both the simple average model as well as the subjective confidence weighted model.

Besides weighting the expertise in the whole crowd, one can also use a different approach: find professionals in a group, weight only them and get rid of the rest. This strategy leads to a compromise between quantity and quality of the crowd. One of the models based on this approach is the contribution weighted model (hereafter CWM) designed by Budescu and Chen (2014). The authors firstly calculated expertise score based on the subject's contribution to the aggregated performance of the crowd (measured by a quadratic scoring rule). Subsequently, they chose only people whose score was larger than zero and normalized this score to create weights that were used for new forecasts. In their paper, static and dynamic version of the CWM (individual contribution updated regularly) was used. Both variants outperformed the simple average model as well as the absolute performance weighted model, additionally, the dynamic version in general outperformed the static version. Consequently, Chen et al. (2016) tested the robustness and cost-effectiveness of the CWM and analyzed the effect of training and collaborative environment on the crowd's expertise. By using longitudinal forecasting study, the Good Judgment Project, they not only supported findings from Budescu and Chen (2014) but also proved that providing access to information via training in probabilistic forecasting or allowing for collaboration in a team results in even greater accuracy of the CWM. As for the robustness of the model, forecasts made in the beginning, during or at the end of the forecasting period yielded stable results and still outperformed the simple average and the absolute performance weighted model. Importantly, one of the biggest advantages of the CWM lies in the fact that even with lower number of judges and limited availability of past performance, it works well. Furthermore, it can identify dishonest forecasters and assign to them lower contribution so that their impact on the overall performance of the crowd is minimized. Mannes et al. (2014) also proposed that selecting a smaller group within a larger set delivers better results. Inspired by Cooke's method, their selection criteria was based on how accurate judges were in recent predictions on unemployment, inflation etc. Their accuracy was measured by the absolute error of the estimate: the smaller the error the better the accuracy. The results showed that the model using only the pre-selected subjects scored better than both the individual estimate and the simple average. Similarly, Hill and Ready-Campbell (2011) used the past performance approach to com-

pare crowd's stock picking decisions with the S&P 500 and concluded that the smaller group picking ability was better than both the larger crowd as well as the S&P 500 performance. To sum up, sometimes it is better to focus only on a subsample of the subjects who have a higher level of knowledge because this approach can deliver better accuracy of answers and is most of the times cheaper.

In general, deciding about what aggregation method to choose depends on the type of task, amount of time available, the composition of the crowd and on budget constraints. Naturally, understanding which approach is the most efficient leads to better decision-making in personal life as well as in economic, political and other spheres. As shown above, the performance of several models has already been compared and evaluated. However, there are still many cases that have not been a subject of a scientific scrutiny yet and that is also where this thesis is directed. In the following sections, my main focus will be on comparing the performance of three models: the simple average (SA), the SPA algorithm, and the CWM. My motivation on this matter is as follows: Prelec et al. (2017) showed in their paper that the SPA performed better than the subjective-confidence weighted model both in easy and tricky questions. However, Budescu and Chen (2014) used subjective confidence in building their CWM and concluded that their model yielded very good results and similarly as the SPA performed well in cases when the majority of the crowd was wrong. Therefore, the question arises quite naturally whether the CWM could improve the subjective confidence weighted model used in the paper from Prelec et al. (2017) and *outperform* the SPA, namely in the domain of general knowledge. The SA is in this case used as a basic aggregation model for comparison. As it is expected to underperform both the CWM and the SPA, it serves as a sort of control that the models were properly constructed. Based on the explanation above, I formulate the null hypotheses of this paper that are going to be tested in the model later on in the following way:

H01: The simple average aggregation method performs equally as the CWM and the SPA in the general knowledge domain.
H02: The CWM performs equally as the SPA in the general knowledge domain.

Chapter 3

Methodology

This chapter will first describe the two main aggregation algorithms used in this thesis, namely the contribution weighted model (CWM) and the surprisingly popular answer algorithm (SPA). Importantly, the explanation of the first method will be based on Budescu and Chen (2014) and Chen et al. (2016), and of the second method on Prelec et al. (2013) (working paper) and Prelec et al. (2017). The second part of the methodology is devoted to experimental design and data description. Finally, I will conclude this chapter with a focus on data analysis including the explanation of the tests used for performance comparison, robustness checks etc.

3.1 Contribution Weighted Model

The CWM's main focus is on finding experts in the crowd based on their relative contribution to the aggregated performance of the group. Each individual's contribution is calculated as a difference between the outcome of the group with and without this individual. Once all the subjects' contributions are known, the CWM is constructed by using only those people whose score is positive.

Let S be a group of subjects who provide forecasts about various questions or events, where the given event/question is $i = 1, \ldots, N$.[1] Consider F_{si} to be a forecast of subject s about event i and F_{Si} to be crowd's aggregated forecast about the event i, where $F_{Si} = A(F_{si})$ and $A(\bullet)$ is a function used for aggregation (e.g. mean). In order to calculate each subject's contribution to the crowd, one has to exclude him or her from the group (one at a time)

[1]Note that there is no need to assume that each subject provides forecasts on the same set of questions.

and recalculate the collective forecast again. This new forecast of the smaller group is specified as $F_{(S-s)i}$. Subsequently, it is necessary to mathematically evaluate the forecasting accuracy of the whole and reduced group via a merit function $f(\bullet)$:[2] $M_{Si} = f(F_{Si})$ (performance of the whole group) and $M_{(S-s)i} = f(F_{(S-s)i})$ (performance of the reduced group), where $s = 1, ..., S$. The relative contribution of subject s to the group's forecast of event i is then denoted as C_{si}, and expresses the difference between the performance of the whole group and the reduced group: $C_{si} = M_{Si} - M_{(S-s)i}$.

If a quadratic scoring rule is used as a merit function then the evaluation of the crowd's performance for event i looks like:

$$Q_i = a + b \sum_{r=1}^{R_i} (o_{ir} - m_{ir})^2, \tag{3.1}$$

where R_i is the number of classes used in predicting event i (e.g. if binary question then 2 classes), $r = 1, ..., R_i$ represents the possible outcome of each event i, m_{ir} stands for the aggregated mean forecast of the crowd about event i (noted above as F_{Si}), o_{ir} is an indicator whether an outcome i_r happens or not (1 = occurs, 0 does not occur) and a and b are constants with values 100 and -50, respectively. Based on a and b, the quadratic score ranges between 0 and 100 where the lower bound means the poorest performance and the upper bound the best possible result. The relative contribution of subject s to the crowd's performance in prediction of the event is then calculated as a difference between the whole and reduced crowd's quadratic score:

$$C_{is} = Q_i - Q_i^{-s} \tag{3.2}$$

For the logarithmic scoring rule, the performance of the crowd is evaluated based on $L_i(r) = ln(r_i)$ where r is an aggregated mean forecast assigned to the outcome that eventually occurs (Bickel 2007). The individual's relative contribution is then calculated as $C_{is} = L_i - L_i^{-s}$.

C_{is} can be either positive or negative meaning that the crowd has either worse or better performance without the target subject. There is also a possibility that the contribution is equal to zero indicating that the crowd's performance has not been affected by the individual.

After calculating individual contributions for each event i, the next step is

[2]This can be done only after the real outcome of the event is known.

to define subject's *total* contribution to the crowd's performance by averaging across all N_s questions that were responded by the subject s: $C_s = \sum_{s=1}^{N_s} \frac{C_{si}}{N_s}$. The concrete form of subject's average contribution with the quadratic or logarithmic scoring rule looks like:

$$C_s = \sum_{s=1}^{N_s} \frac{(Q_i - Q_i^{-s})}{N_s}, \quad C_s = \sum_{s=1}^{N_s} \frac{(L_i - L_i^{-s})}{N_s} \tag{3.3}$$

Once the total contributions of all the subjects in the group are known, the CWM can be constructed. The strategy is to take only those people, whose average contribution is higher than certain threshold, normalize this contribution and create weights scaled such that $w_s = 0$ if $C_s \leq \epsilon$, and $w_s = (\frac{C_s}{\sum_{s=\epsilon}^{S} C_s})$ if $C_s > \epsilon$. The ϵ is usually set to zero, which is also the case in this paper. However, one can also choose to use a stricter bound to include only best performing subjects in the model ($\epsilon > 0$). Finally, the calculated weights are used for forecasting a new set of questions/events.

Importantly, Budescu and Chen (2014) pointed out in their Study 1 that the CWM has two parts: firstly, it identifies experts according to their positive contribution, which accounts for 60% of the success, and secondly, it weights these people based on how high their positive contribution is (40%). Therefore, finding experts with superior knowledge is more important than weighting them. Moreover, the authors of the model noted that in order to make sure that the weights are not based on subject's luck in questions, at least 10 events should be used to calculate the individual's average contribution. However, in the ideal case, the proper measure would be based on about 25 - 30 events happening in several different points in time. Importantly, the predicted events should come from some specific domain, such as geography or politics. Otherwise, subjects might get a positive score from spheres where they are good at and a negative score for the rest of the predictions and therefore, there would be no 'absolute' professional for the whole set of questions. In such situations, the weights denoting the experts would lose sense.

The methodology presented above concerned a static version of the CWM, however, there exists also a dynamic form (based on similar principles) that includes time perspective and yields even better results. The logic is as follows: firstly, the initial weights are, similarly as in the static version, constructed based on some basic set of events/questions. Secondly, the weights are updated every time the outcome of a predicted event is observed. In this manner,

subjects who were once excluded from the model because of their low contributions can get in the model again with time and vice versa. Even though the dynamic version performs better than the static one, I use the latter form in this paper because of unavoidable time constraints.

3.2 Surprisingly Popular Answer

Prelec et al. (2017) developed the SPA algorithm in order to overcome the drawbacks of the simple average and the subjective confidence weights. Namely, one of the biggest disadvantages of taking the mean is that if there are too many people who do not know the correct answer, the result will be wrong. The subjective confidence weights improve the SA a bit as they usually better account for extra knowledge, however, in many cases people might be confident about a wrong answer, which results again in inefficient outcome. For example, imagine a statement such as: "The tomato is a vegetable". In such a situation, people who are not educated in botanics (thus the majority) will confidently endorse the sentence because they will recall eating the tomato in their spring salads along with other vegetables. However, their conjecture would be false as the tomato is botanically classified as a fruit. Consequently, both the simple average as well as the confidence weights will fail. The SPA, however, has an ability to identify the correct answer in this case. The algorithm asks subjects to express their opinion on the statement and predict the distribution of other people's answers. Subsequently, it chooses the response that is more popular than originally expected, which is also the correct answer. The principle based on which the method works will be described in the following paragraphs.

In order to understand the SPA, imagine that there are two possible worlds: the real world in which the tomato is indeed a fruit and the counterfactual one, in which it is a vegetable. Naturally, we can assume that in the actual world less people would say that the tomato is a vegetable than in the counterfactual, however, in both cases, the majority would be wrong. The authors of the method help to explain this by the toss of a two-sided biased coin: in the real world the coin (people) says 'vegetable' 70% of the time and in the counterfactual world 95% of the time. Rational respondents assume the existence of the two worlds, moreover they understand what the biases of the coin are. However, they do not know which of the two worlds is the actual one (the world with 75% or 90%). As a result, they would predict that the proportion of people endorsing 'vegetable' will fall somewhere between the lower and up-

per bound, let us say 81%. In the real world, however, the number of votes for vegetable will be closer to 70%. Therefore, one can identify the actual world and thus the correct answer as the one where the frequency of the answer is higher than predicted by the respondents, which is, in our case, the 'fruit'. This reasoning explains why it is important to care not only about the sole answer to a question but also about subject's estimation of other people's answers to that question.

Mathematical intuition is as follows: a question has m possible answers where each of these answers stands for one world. Uncertainty about which world is the real one is represented by a random variable that takes on values from $\{a_1, ..., a_m\}$ (note that each of these worlds contains an n-sided coin). Further, the answer (vote) V^r of respondent r is based on signal T^r, which can be expressed as a random variable taking on values from the set $\{t_1, ..., t_n\}$. Therefore, the vote is formally denoted as $V^r = V(T^r)$ and $V^r \in \{v_1, ..., v_m\}$. Consequently, there are three necessary assumptions that need to hold. First, the signal of each individual is independent of other individuals' signals and all the signals are identically distributed with probabilities $p(t_r|a_i)$. Second, each subject gives an honest response according to the signal he or she gets. Finally, people with different answers are provided with different signals and people with the same answer use the same signal.

Ideally, the subjects are aware about the worlds' probability distribution $p(t_r, a_i)$. Yet, they do not know which of the worlds is the actual one and the distribution of obtained signals. In other words, the coins' availability and their biases are clear, however, nobody knows which of them is actually tossed.

Finally, the authors explain that subjects hold two different kinds of beliefs: the first kind indicates the uncertainty about the correct answer and is denoted by the posterior probabilities $p(a_i|t_k)$. The second kind of beliefs represents the uncertainty about the distribution of other people's signals $p(t_j|t_k)$ and can be obtained from asking subjects to predict the other subjects' answers. Ideally, the respondents would use Bayes' rule to calculate the distribution of other signals across the two worlds: $p(t_j|t_k) = \sum_i p(t_j|a_i)p(a_i|t_k)p(a_i)$, where $p(a_i)$ is a common prior and is defined as shared information among all subjects, for example: the tomato is served along with other vegetables in the salad.

If for each available signal, we define $p(a_i|t_k)$ as a probability that the signal assigns to the correct answer (actual world), then the best answer is the one that yields the highest possible probability. In order to calculate this term directly one would need to find out how the probabilities are distributed over

potential states of the world for the *unknown* value of $i = i^*$, which is hard to achieve. In this paper, I use questions that have binary options 'true' and 'false', therefore $m = 2$ and $n = 2$ (two worlds each with one two-sided coin). For such a case, the following trick suggested first by Prelec (2013) is used to elicit the true answer:

$$\frac{p(a_{i^*}|t_1)}{p(a_{i^*}|t_2)} = \frac{p(t_2|t_1)}{p(t_1|t_2)} \frac{p(t_1|a_i)}{p(t_2|a_i)} \tag{3.4}$$

If the ratio on the left side is larger than one then $p(a_{i^*}|t_1) > p(a_{i^*}|t_2)$ and 1 is the correct answer, if it is smaller than one then $p(a_{i^*}|t_1) < p(a_{i^*}|t_2)$ and 2 is the correct answer.

To relate the equation (3.4) to the SPA explanation, let us take again the example with the tomato where 1 means that the answer is a fruit and 2 that it is a vegetable. The second term on the right side of the equation expresses the ratio of the democratic votes for fruit versus vegetable. The divisor is going to be larger than dividend as most of the people will think that the tomato is a vegetable, thus the whole term will be smaller than one. In the first term on the right side, the dividend expresses prediction of 'fruit' voters about the frequency of 'vegetable' voters. As the shallow and commonly shared information is that the tomato is a vegetable, people who voted for fruit will expect most people to say 'vegetable'. On the other hand, those who would endorse vegetable will predict that the others will agree with them (divisor). Therefore the whole term will be larger than one, and will possibly overcome the second term on the right side. Consequently, $p(a_{i^*}|t_1) > p(a_{i^*}|t_2)$, thus 1 will be the correct answer. The definition of the SPA is met as the average of the votes for the correct answer will be underestimated (both 'fruit' and 'vegetable' voters will predict vegetable as being more frequent), therefore, the 'fruit' will be surprisingly popular answer.

Prelec et al. (2013) explained that there is also an alternative approach how to get $p(a_{i^*}|t_k)$, namely by using Bayesian Truth Serum (BTS) score. The correct answer k is then the one that maximizes the following term:

$$\bar{u_k} = \frac{1}{n\overline{x}_k} \sum_s x_k^s u^s \tag{3.5}$$

Since $\lim_{n->\infty} \bar{u_k} = logPr[a = i|T^r = k] + C$, (3.5) is clearly just a different way how to express $p(a_{i^*}|t_k)$ (Prelec and Seung 2006). In the equation u^s is the

BTS score of individual respondent s and is defined as:

$$u^s = \sum_{k=1}^{m} x_k^s log \frac{\overline{x_k}}{\overline{y_k}} - \sum_{j=1}^{m} \overline{x_j} log \frac{\overline{x_j}}{y_j^s} \tag{3.6}$$

In (3.6), x_k^s takes on values from set $\{0, 1\}$ and indicates whether subject s endorses the answer k or not, $\bar{x_k}$ is a simple average: $\overline{x}_k = \frac{\sum_s X_k^s}{n}$, y_j^s is a prediction of subject s about the distribution of answer j and $log\bar{y_j} = n^{-1} \sum_s \log y_j^s$. The first term in the equation refers to so called information score and is the same for all the subjects who endorse the same answer. The second term refers to the prediction score and serves as a penalty for incorrect estimation of other people's answers and therefore, the better is the prediction the lower is the deduction. Importantly, Prelec et al. (2013) pointed out that instead of using the SPA algorithm, one can also take answers of subjects with the highest BTS score and obtain similarly accurate results. This is also the strategy I am applying in my thesis where one of the approaches is to identify experts based not only on the CWM score but also and on the BTS score.

3.3 Experimental Design and Data Description

In order to compare the performance of the SA, CWM and SPA, I generated an anonymous quiz both in paper and online form. In the first case, the answers were collected under my supervision in the student canteen at the campus of Erasmus University Rotterdam. In the second case, I used the research software Qualtrics and distributed the questions via social websites to (mostly) university students across the whole Europe. Note that contrary to the paper form, there was no control in the online version. The two different ways of data collection were chosen to check if subjects without supervision had a different answering pattern, which might indicate cheating. Based on the nature of the quiz, there were two main reasons for swindling. First, the questions required only a very basic level of understanding of the topic, therefore, it would feel shameful for subjects to not answer them correctly. Second, respondents knew that they would be scored and compared with others (even though only anonymously), consequently, they wanted to achieve the best result possible.

The quiz itself consisted of three parts. Firstly, there was a sheet with instructions which also included a sample question and a description of two scores that would be assigned to each individual based on his or her answers.

The next section consisted of five questions that were targeting gender, age, education, nationality as well as the number of hours each person watched or read news per week. Finally, the core part asked 35 general knowledge questions from 5 different domains (each domain consisting of 7 questions): geography, science versus fiction, inventions, word cultural heritage, and world 2016. This division was done in order to make sure that subjects would not get bored and stop paying attention, however, note that all the spheres were quite substantially overlapping. In fact, a significant majority of the questions could be, in a broader sense, considered as a part of geography and inventions domain. This particular design satisfied the CWM requirement of choosing tasks from a narrow sphere in order to efficiently find real experts in the field.

Importantly, to avoid bias towards using too 'tricky' or easy topics that would favour one of the compared models more than the other, the questions were randomly selected from the online quiz website: `www.quipoquiz.com`. In the paper form, ten versions with a different order of the questions were used to account for the order effect and also for the risk that subjects would lose their attention towards the end of the quiz. These versions were generated randomly. In the online form, the change in order was done automatically by the Qualtrics software.

Each question in the core part of the quiz consisted of a statement and three subparts. In subpart a), subjects should decide if they agree or disagree with the statement. In subpart b), they had to express their subjective confidence about the answer in a) on the interval scale from 50% to 100%. This interval was chosen over the interval from 0% to 100% to make sure that respondents did not give confidence lower than 50%, which would lead to inconsistency. Finally, subpart c) asked participants to estimate, based on their subjective belief, the proportion of people who said that the original statement was true (on the interval scale from 0% to 100%). To give a concrete example, one of the questions in the world 2016 domain looked as follows:

The deadliest earthquake of 2016 took place in Italy.

(a) Your answer: True/False

(b) What do you think is the prob. that your answer is correct:

(c) Think about other people's beliefs and predict the percentage of people who said the answer was 'True':

Note that subparts a) and b) collected information for the construction of the CWM. In the original paper from Budescu and Chen (2014), subjects were making predictions about unknown outcomes in the *future*. However, due to time constraints, I targeted on questions to which the answers were already known. As for subpart c), it collected, along with subpart a), the data necessary for the SPA algorithm.

Since the quiz was quite time consuming, it was important to provide motivation for honest responding and attention. In this case, the monetary rewards were inadequate because the evaluation of responses took too much time to provide immediate remuneration, quiz was anonymous and the budget for the research was low. Therefore, I introduced a competitive element so that participants could compare their performance in the quiz with others. This incentivization worked as follows: all general knowledge answers were evaluated based on two criteria: 1) the objective accuracy score and 2) the prediction score. The first score evaluated subparts a) and b) and was based on the Brier scoring rule which has the following form: $S(p,1) = 2 - (1-p)^2$ if a predicted outcome happens and $S(p,0) = 2 - (p)^2$ if the outcome does not happen. The score was constructed in a way that if a participant was confident about a correct answer, she got bonus points (maximum 25), however, if she gave high confidence about a wrong answer, the penalty was in absolute terms by far stricter than the bonus (the lowest possible score = -75). The prediction score depended, on the other hand, on the accuracy of predictions of other people's answers in subpart c) and was scaled in the same way as the objective accuracy score (max = 25, min = -75).

Table 3.1: Individual Scores

Rank	**Nickname**	**Objective accuracy score**	**Prediction score**	**Final score**
1	Michal	419	778	1197
2	Vojta	412.7	783	1195.7
3	SaltCaramel	378.5	792	1170.5
4	Nils	377.8	757	1134.8
5	Vrut	350.1	774	1124.1
98	Cesar	-546.8	677	130.3
99	Wenjun Fu	-517	643	126
100	Saltic	-477.5	595	117.5
101	Antonio	-417.5	518	100.5
102	Paula	-471.8	367	-104.8

For each subject, the two scores were summed up to create a final score based on which a rank was assigned. In the beginning of the quiz, the re-

spondents could voluntarily write down their name or nickname so that they could check the correct answers afterwards and see how they performed on goo.gl/ovkvFO. The top 5 highest final scores were all above 1120 threshold, whereas the lowest 5 scores were all below 135. Furthermore, people with high objective accuracy score usually had a high prediction score and vice versa, for more information see Table 3.1 (or check the link).

The collection of the data took 8 days. In total, 102 participants filled in the quiz, 51 online and 51 in the paper form. The gender of the participants was quite balanced: 49 women and 53 men, the average age was 23 years. Most of the people came from the Netherlands, had finished their bachelor studies and spent on average 5.7 hours by reading or watching news. Some subjects did not want to fill in the demographic questions so that some observations were missing for the particular characteristics, see Table 3.2.

Table 3.2: Summary Statistics

Variable	**Observations**	**Mean**	**Standard Dev.**	**Min**	**Max**
Gender	102	1.52	0.50	1	2
Age	100	23.10	2.93	18	35
Education	100	2.25	0.96	1	3
Nationality	79	4.91	2.97	1	13
News - hours/week	98	5.74	4.37	0	30

The next step after collecting the data was the analysis itself. Therefore, in the following section, I will provide an overview of the procedures alongside with parametric and non-parametric tests that were used to evaluate the performance of the SA, CWM and SPA as well as to run a supplementary analysis concerning the BTS score, possible cheating in online and paper form, and robustness checks.

3.4 Data Analysis

Based on the CWM theory explained in Section 3.1, I applied the quadratic scoring rule to derive each individual's contribution for each of the 35 general knowledge questions. Subsequently, I randomly chose ten of these questions and calculated every subject's total contribution for this subset in order to obtain individual weights. People with zero or negative average contribution in the 10 questions got a zero weight ($w_s = 0$) and those with a positive average contribution were assigned with a positive weight normalized according

to the size of their contribution ($w_s > 0$). Finally, the weights were used for determining the answers of the remaining 25 quiz questions. As for the SPA, I applied formula (3.4) which was introduced in Section 3.2 and for the SA calculation I used the following equation: $\overline{x}_k = \frac{\sum_s X_k^s}{n}$. Finally, to provide more insights into the analysis, I also included an algorithm based on the BTS score in the comparison as it is closely related to the SPA algorithm (see Section 3.2). In this case, I took the simple average of the subjects' answers from the a) part of the questions whose BTS score was positive. To summarize, all the algorithms were applied to each of the 35 questions in the quiz. However, only 25 questions were eventually used for the analysis because the rest determined the weights of the CWM. If a model delivered a correct answer to a question then this answer was coded as 1 and when incorrect, it was coded as 0 and therefore, the obtained data was binary with $n = 25$ for each of the algorithms.

To assess whether the number of correct answers was statistically different from chance for each of the models, I applied a one-sample binomial test. Its outcome was a p-value which if close to zero meant that the number of correct answers was too high to be random. Subsequently, three different tools were used to compare the performance of the three models (plus the BTS algorithm): McNemar test which is the most relevant for the analysis as it is meant for two samples with matched pair nominal data (thus our situation), the matched pair sign test that was used by Prelec (2017) (I apply it only for six crowds - see the explanation below), and the t-test. Importantly, the first two tests are non-parametric so that they do not need to assume normality and can be applied to small samples, however, they are weaker than the parametric t-test.

Firstly, McNemar test compares the number of cases when one of the two samples performs well and the second one poorly and vice versa. If the difference in this performance is statistically significant then one model is better than the other. Secondly, a matched pair sign test is generally applied to paired data on at least ordinal scale, however, application to nominal data is, with certain limitations, also possible. The procedure is as follows: for each question, the answer from the first model is subtracted from the answer of the second model. As a result, a new sample with the possible values -1,0,1 (since the data is binary) is generated. Subsequently, the binomial test is applied on this new sample to determine if there is a difference between the number of positive and negative signs. If yes, the performance of the models is statistically different. Unfortunately, the biggest disadvantage of the sign test is that for binary data, many times the difference between the two models is equal to

zero which the binomial test does not account for and therefore, many observations are lost. That is also the main reason why I used this test only for 6 crowds (see the explanation below) and not for the original crowd consisting of 102 subjects. Finally, the t-test was applied on the difference between two models. It assumes normality, which is for the data set with the values -1,0,1 hard to achieve. Nevertheless, with sufficiently large sample size, it is possible to consider the normal distribution as an approximation of the binomial distribution. In this study, n was quite low (25), which made the applicability questionable. However, the test was used anyway since it could provide some additional insights to the analysis.

Besides using binary codification to evaluate the performance of the models (1 = correct answer, 0 = incorrect answer), one can also rate models' accuracy by F1 score and Matthews correlation coefficient as Prelec et al. (2017) did. Since these tools are used mainly to visually depict the difference in the performance but do not cover significance, I applied them only as a supplementary measure of the models' performance and focused mostly on the parametric and non-parametric tests mentioned above.

Apart from the core analysis, the following topics were covered: first, I focused in a greater detail on the comparison of the BTS and CWM algorithms. Namely, how would the performance of the two models change in comparison with the SPA and SA if I took only 30 or 20 people with the best CWM or BTS scores and did not use the remaining subjects. As Prelec et al. (2013) explained, employing only participants with the highest BTS scores might deliver better results than the SPA. Additionally, Chen et al. (2016) claimed that the CWM with 20 subjects gives almost the same outcome as the one with 50. Therefore, the analysis of the two algorithms with less people naturally made sense.

Second, I examined if there was any relation between the BTS score and the CWM individual score. As both of them award possession of some advanced knowledge, a positive correlation would not be surprising. Moreover, the BTS score serves as a basis for the SPA algorithm (Prelec and Seung 2006) and therefore, analyzing its connection to the CWM might provide some additional insights into the performance comparison of the CWM and SPA in general. In order to dive deeper, I used the CWM score based not only on the quadratic scoring rule but also on the logarithmic scoring rule in this part to see if there would be a different relationship when both of the scores consist from natural logarithms. A regression was used to conduct this analysis in order to account for the effect of demographic factors including age, educations and nationality.

The general equation looked as follows:

$$CWM = \beta_0 + \beta_1 BTS + \beta_2 gender + \beta_3 age + \beta_4 educ + \beta_5 nation + \beta_6 news + u \tag{3.7}$$

Subsequently, as already mentioned in Section 3.3, I analyzed if the subjects in the online version had a different answering pattern than the subjects in the paper form, which would indicate cheating. On that purpose I ran a Fisher exact test for two independent samples ($n_1 = 1785, n_2 = 1785$) to find out if there is a significant difference between the number of correct answers for the two versions.

To see if there would be any performance difference between the SPA, CWM and SA for larger sample size (more questions included), I randomly split the crowd of 102 subjects into six smaller groups (each with 17 subjects), and recalculated the CWM, SPA and SA. As a result, I got six times more observations (150) compared to that of the original crowd. Therefore, I could use the matched pair sign test and still have enough observations even if zeroes were omitted. Moreover, the t-test became more relevant because the normal distribution was better approximated due to higher n. Note that since each of the 6 crowds gave an answer to the same set of 25 questions, the observations became dependent. The point is that every question appeared in the sample 6 times. If the question was easy, most of the crowds answered it correctly, if it was difficult, the crowds failed. Thus the six responses (observations) to this question became correlated in the sample. As both parametric and nonparametric tests assume independence of observations, this becomes a major limitation of the large sample analysis. Importantly, even though the crowds consisting of 17 people might seem to be too small, based on Wagner and Suh (2014), the wisdom concept can efficiently work already with less than 10 people. Moreover, a larger crowd does not necessarily provide better results because the group knowledge includes diminishing marginal returns from the sample size. As a result, over a certain threshold the number of people does not enhance the performance anymore. Finally, note that this time the BTS algorithm was omitted, first, because the analysis with 6 crowds is only supplementary and the BTS is not the main focus of this paper, second, due to inevitable time constraints.

Furthermore, I also tested how lowering or increasing the number of weighting questions changes the results. Budescu and Chen (2014) suggested that the lowest possible amount is 10 (otherwise a good score of an individual might be

just due to chance). Moreover, they noted that the more questions are used, the better the CWM works. In this paper, versions with 8 to 13 weighting questions were used on the crowd to analyze if there is any significant difference in the outcome.

Finally, robustness checks were done thorough the paper. In the original crowd with 102 subjects, I randomly chose 20 different sets of 10 weighting questions to see if the performance changed with different weights. As for the six groups of smaller crowds, I chose two approaches. First, for one set of 10 weighting questions, subjects were randomly divided into the six groups twenty times, each time the performance was compared. Second, for one randomly chosen crowd, I changed the set of 10 weighting questions 20 times and analyzed the performance.

Chapter 4

Results and Discussion

4.1 Results

In this chapter, I will outline the main results based on the procedures discussed in Section 3.4. Moreover, I will focus on the limitations of the study alongside with suggestions for further research.

First, the outcomes of the binomial test, which was used to assess if the number of correct answers delivered by each of the models was statistically different from chance, are summarized in Table 4.1. Note that to check for robustness, 20 different sets of 10 weighting questions were used in the analysis. Every time when the p-value of the test was less than 0.1, the case was noted as 'significant'.

Table 4.1: Binomial Test

Algorithm	Significant Cases	Insignificant Cases
CWM	20	0
SPA	20	0
SA	19	1
BTS	20	0

Based on the figures for the CWM and SPA, the high number of correct answers could not have been coincidence for any of the 20 versions with different weights. The same holds for the BTS. Additionally, there was only once an insignificant difference between correct and incorrect answers for the SA. Therefore, using the above mentioned algorithms to obtain the highest amount of correct answers possible is better than determining the answers randomly.

The next step was to compare the performance of the models. This was done again for 20 different sets of weighting questions to check for robustness. As all

the applied tests can analyze only two samples at a time, there were six possible combinations for the comparison: CWM and SPA, CWM and SA, SPA and SA, CWM and BTS, SPA and BTS, and BTS and SA. Firstly, McNemar test was used for the analysis. Table 4.2 shows that for all the paired comparisons, there was no significant difference between the models except for three cases in the CWM/SA.

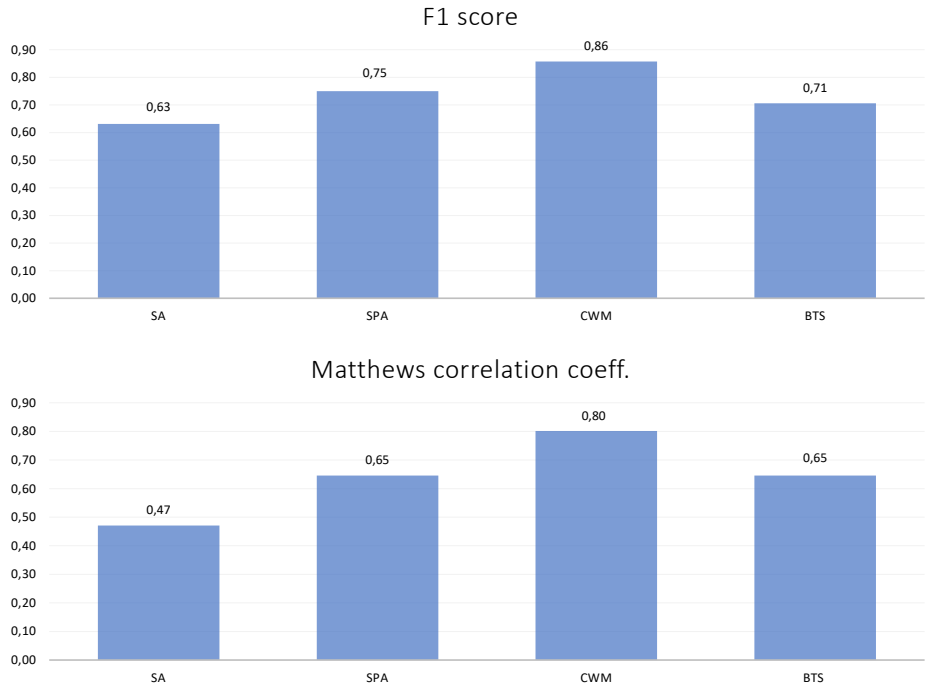
Table 4.2: Tests for Performance Comparison

	McNemar		**T-test**	
Difference:	Significant	Insignificant	Significant	Insignificant
CWM/SPA	0	20	3	17
CWM/SA	3	17	5	15
CWM/BTS	0	20	10	10
SPA/SA	0	20	7	13
SPA/BTS	0	20	0	20
BTS/SA	0	20	6	14

When the parametric t-test was used, the number of significant differences increased for all the model comparisons. Nevertheless, the ratio of significant versus insignificant cases was never distinct enough to conclude that one of the models performed better than the other (even for the CWM and BTS, where there were 10 significant differences). To sum up, I did not find any significant difference in performance between any of the models when applying the two mentioned tests. Importantly, the different results from each of the tests are caused by different nature of these tests. Namely, the t-test requires normal distribution, which is with 25 observations and nominal data hard to achieve and therefore, its outcomes need to be taken with caution. Finally, as for the performance in significant cases of the tests: the CWM always outperformed the remaining models, the SPA was almost every time better than both the BTS and SA, and finally, the BTS delivered better results than the SA in the large number of the cases.

To illustrate the comparison of the models graphically and thus provide a better overview, the graphs in Figure 4.1 depict F1 score and Matthews correlation coefficient for one randomly chosen set of weighting questions (set number 3). The picture shows that the CMW scored the best with F1-score and Matthews correlation coefficient equal to 0.86 and 0.80, respectively. The second one was the SPA followed by the BTS and ending with the SA algorithm. Based on the small distance among the numbers, there was no distinct difference between the models both for the F1 score and MC coefficient.

Figure 4.1: F1 Score and Matthews Correlation Coeff.



Source: Author's computations.

After the direct comparison of the models, the next step was to check how the performance of the CWM and BTS algorithms changed when only subjects with the best scores were used. On that purpose, I randomly chose two sets of weighting questions and analyzed whether the number of correct answers changed as the subjects with lower scores were step by step omitted from the two models. The results are summarized in Table 4.3, where the labels 'correct' and 'incorrect' stand for the number of questions answered correctly/incorrectly. Apparently, taking only top 10 subjects does not seem to deliver a significant difference in the performance from taking top 40 subjects, both for the CWM and BTS algorithm. This finding is quite surprising, however, it could be explained by several factors. The first reason might be the nature of the questions. Namely, in some of the tasks most of the people were very sure about a certain answer and in other questions most of the subjects had no clue. However, there was a relatively low number of questions where the proportion of people being correct or incorrect would be mixed. Simply, the crowd either knew or did not know. Therefore, taking less respondents did not make a bigger difference in the number of correct answers. Moreover, the analysis was limited due to small sample size so that making any conclusions turned

out to be rather difficult. Despite the reasons mentioned above, this finding gave us some insights about the cost effectiveness, namely, it is not necessary to use a large number of respondents for the CWM and BTS calculations.

Table 4.3: Performance with Less Subjects

Weights:	7;8;9;11;13;19;26;27;32;35		2;10;14;15;16;18;19;24;26;35	
Model	Correct	Incorrect	Correct	Incorrect
SA	17	8	19	6
SPA	20	5	22	3
CWM51	21	4	22	3
CWM40	21	4	22	3
CWM30	21	4	21	4
CWM20	22	3	21	4
CWM10	19	6	22	3
BTS51	19	6	21	4
BTS40	21	4	23	2
BTS30	21	4	22	3
BTS20	21	4	23	2
BTS10	18	7	23	2

Looking at the insignificant difference in the performance of the CWM, SPA and SA, it is important to understand the reasons for such an outcome and therefore, I provide three possible explanations. First, an important factor for an indifference between the SPA and CWM might be a positive relationship between the CWM score and the BTS score since as explained in Prelec and Seung (2006), the SPA is based on the BTS. In order to test this conjecture, I regressed both the quadratic and logarithmic version of the CWM score on the BTS score in total 20 times (to account for different weights) as explained in Section 3.4. In the version with the quadratic scoring rule, 19 out of 20 regressions had a positive and statistically significant BTS coefficient with the average size of 0.222 and a standard deviation of 0.071 (in 17 cases, the statistical significance was 1%). Similarly, with the logarithmic scoring rule, the BTS coefficient was again positive and statistically significant in 19 out of 20 regressions with 1% significance in 17 regressions. The average size of the coefficient was 0.0047 with a standard deviation of 0.001625. To sum up, both the quadratic and logarithmic score had a positive and statistically significant relationship with the BTS score, which at least partially explains why the difference between the SPA, BTS and CWM algorithms was insignificant. Moreover, there was no difference in the statistical significance of the BTS score when the logarithmic and quadratic version of the CWM score was used. Importantly, none of the control variables including age and education

had a robust significance in the regressions. To provide basic understanding, I show 5 out of the 20 regressions for each score in Table 4.4 (based on sets of weighting questions: 1, 5, 11, 15, 20). The complete tables can be found in Appendix.[1]

Table 4.4: Regression of the CWM Score on the BTS Score

Quadr. s. r.	(1) CWMScore	(5) CWMScore	(11) CWMScore	(15) CWMScore	(20) CWMScore
BTSScore	0.198***	0.221***	0.295***	0.295***	0.370***
	(0.0506)	(0.0648)	(0.0662)	(0.0786)	(0.0660)
Gender_dummy	-0.00452	0.00648	0.0282*	0.0158	0.0226
	(0.0144)	(0.0155)	(0.0158)	(0.0160)	(0.0151)
Education_cat	0.00759	0.0154	0.00559	0.0127	0.00536
	(0.00857)	(0.00989)	(0.00922)	(0.00988)	(0.00997)
Nationality_cat	0.00336	0.000667	0.00767**	0.00202	0.00448
	(0.00305)	(0.00354)	(0.00363)	(0.00407)	(0.00278)
Age	-0.00107	-0.00134	0.00109	-0.00773**	0.00310
	(0.00303)	(0.00316)	(0.00317)	(0.00330)	(0.00335)
Newshours	0.000662	-0.000680	0.000569	9.41e-06	0.00140
	(0.00193)	(0.00176)	(0.00200)	(0.00178)	(0.00150)
Constant	-0.000706	-0.00627	-0.118	0.121	-0.146*
	(0.0690)	(0.0696)	(0.0774)	(0.0740)	(0.0770)
Observations	77	77	77	77	77
R-squared	0.203	0.216	0.382	0.335	0.511
Logarithm. s. r.	CWMScore	CWMScore	CWMScore	CWMScore	CWMScore
BTSScore	0.00470***	0.00515***	0.00656***	0.00626***	0.00784***
	(0.00104)	(0.00138)	(0.00133)	(0.00161)	(0.00137)
Gender_dummy	2.73e-05	0.000232	0.000631*	0.000343	0.000485
	(0.000307)	(0.000335)	(0.000324)	(0.000332)	(0.000317)
Education_cat	0.000179	0.000363*	0.000126	0.000273	0.000126
	(0.000179)	(0.000211)	(0.000189)	(0.000205)	(0.000208)
Nationality_cat	7.69e-05	2.26e-05	0.000161**	4.61e-05	9.74e-05*
	(6.44e-05)	(7.51e-05)	(7.40e-05)	(8.46e-05)	(5.78e-05)
Age	-2.15e-05	-2.55e-05	1.70e-05	-0.000157**	6.67e-05
	(6.23e-05)	(6.58e-05)	(6.30e-05)	(6.73e-05)	(6.98e-05)
Newshours	1.27e-05	-1.61e-05	9.54e-06	-4.37e-08	2.69e-05
	(4.21e-05)	(3.85e-05)	(3.99e-05)	(3.78e-05)	(3.15e-05)
Constant	-0.000294	-0.000456	-0.00241	0.00236	-0.00315*
	(0.00142)	(0.00145)	(0.00154)	(0.00151)	(0.00158)
Observations	77	77	77	77	77
R-squared	0.246	0.256	0.417	0.347	0.521

The second reason for the insignificant difference among the models might be that some subjects cheated in the quiz, which would lead to a higher number of positive answers than what would be the case in a natural situation. As a

[1]Since 25 respondents did not state their nationality and education, their answers were omitted from the regression.

result, the advantages of individual models would not be sufficiently materialized. The Fisher exact test for 1184 correct answers out of 1785 in the online form and for 1012 correct answers out of 1785 in the paper version delivered a p-value of 0.00 meaning that the difference in the number of correct answers between the two groups was not random. This finding was further supported by substantially higher average time of filling the quiz in the online form (average time in online version 41 minutes and offline version 26 minutes). Even though, more evidence would be needed to firmly conclude that cheating was present in the online version, clearly the pattern in answering was different between the two groups which might have influenced the performance of the models. [2]

Finally, the third reason might be a small sample size. To see whether the tests would show different results if n was larger, I split the crowd of 102 subjects into 6 smaller crowds, so that I got $n = 150$ and compared the SPA, SA and CWM again. This time I also used the matched pair sign test in the analysis. The results are summarized in Table 4.5:

Table 4.5: Six Crowds - Different Groups and Weights

		20 CROWDS		
Test	Significance	CWM/SPA	CWM/SA	SPA/SA
McNemar	Significant	0	11	18
	Insignificant	20	9	2
Ttest	Significant	1	13	19
	Insignificant	19	7	1
Sign Test	Significant	0	11	18
	Insignificant	20	9	2
		20 WEIGHTS		
Test	Significance	CWM/SPA	CWM/SA	SPA/SA
McNemar	Significant	2	16	15
	Insignificant	18	4	5
Ttest	Significant	2	16	15
	Insignificant	18	4	5
Sign Test	Significant	2	17	15
	Insignificant	18	3	5

The upper part of the table describes the outcomes of the tests where 1 set of weights was used and 20 different versions of 6 crowds were generated. On the other hand, the bottom part shows the results for 1 randomly chosen version of 6 crowds on which 20 different sets of 10 weighting questions were applied. In comparison with the large crowd of 102 subjects, the outcomes of the smaller

[2]The other explanation for the different answering pattern might be that the people filling the quiz online were simply more clever than the people filling the paper form.

crowds delivered significant results. Namely, when one set of weights was used and 6 groups were randomly created 20 times, there was a significant difference between the SPA and SA at least 18 times for all three tests. For one set of 6 crowds and 20 times randomly chosen weighting questions, all tests showed significant difference between the SPA and SA in 15 cases and between the CWM and SA at least in 16 cases (out of 20).

Therefore, there are two main findings that can be derived from the table: First, using 35 questions in the quiz (from which 25 are used for the comparison) is not sufficient to make any conclusions about the difference in the performance of the models. However, if the number increases, like in the 6 crowd case, to 150 (25*6) questions, it is possible to get significant differences. Second, even though the results became more significant for large sample ($n = 150$) than the small sample ($n = 25$), the analysis needs to be taken with caution as the observations with 6 crowds are not independent (see explanation in Section 3.4) so that one of the main assumptions for the tests to work is violated.

To sum up, all three explanations for the insignificant difference among the algorithms - the positive relationship between the BTS and CWM score, the different answering pattern between the online and offline version and the small sample size - have been validated.

Apart from understanding the insignificance, I also studied how the number of questions that were used as weights influenced the number of correct answers delivered. On that purpose, I chose one set of weights and step by step randomly omitted or added tasks such that the lowest possible number was 8 and the highest 13. The results are summarized in Table 4.5.

Table 4.6: Changing the Number of Weighting Questions

	Weights	**CWM**	**SPA**	**SA**	**BTS**
% correct	W - 8	81%	81%	74%	81%
	W - 9	85%	81%	73%	81%
	W - 10	84%	84%	76%	88%
	W - 11	88%	88%	79%	88%
	W - 12	87%	87%	78%	78%
	W - 13	86%	91%	82%	82%

Based on the figures, the correctness did not change significantly with lower or higher number of weighting questions, even though the percentage of correct answers was obviously slightly higher when 13 questions were used rather than 8. The small variance might be explained by the difficulty of the quiz; some questions were easier and a vast majority of the subjects knew the answer to

them while the others were too difficult so that almost nobody had a correct answer. As a consequence, it did not really matter whether I would choose 7, 10, or 13 for weighting because every time, people excelled mostly in easy questions and did not diverge substantially from the rest in more difficult parts.

To summarize, I found out that the SPA and CWM performed significantly better than the SA, but only for the large sample size. Importantly, the performance difference between the SPA and SA was more robust than the CWM and SA as it held both for different weights and crowds. Based on this finding, the first null hypothesis can be rejected, at least for large n. On the contrary, there was no significant difference between the SPA and CWM in any of the cases presented above so that the second null hypothesis cannot be rejected. To conclude, the SPA has a better use than the CWM because both of the models perform similarly, but in the first case, the SPA outperforms the SA more robustly and there is no need for any specific domain and the number of questions.

4.2 Limitations and Further Research

Once the results were summarized, it is necessary to focus on the limitations of the study as well as on suggestions for further research. In this paper, I find 5 main issues that can be tackled in the future.

First, a substantial problem was insufficient incentivization of the subjects that lead to low effort and subsequently to poor results. As mentioned in Section 3.3, money was not used as a motivation because the quiz was anonymous, the evaluation of the answers took too much time to remunerate based on performance and, last but not least, the research budget was low. As a result, I had to introduce the competitive element, hoping it would be enough to boost the subjects' attention and effort. Additionally, I chose questions from the sphere of general knowledge with 5 different domains in order to increase the probability that the subjects would stay vigilant. Despite this effort, many people did not really care about comparing their score and got tired quickly with the quiz. Therefore, I believe that providing a show up fee would deliver better and more relevant outcomes in the future. Namely, by giving, let us say, €10 before the start, I would psychologically commit the subjects to make an effort. As a result, I could afford to give more questions thus increase my sample size and improve the relevance of the tests for the performance comparison. Moreover, with monetary rewards, the sphere of the questions could

become narrower so that the CWM would more precisely choose experts from the crowd, which would probably lead to its better performance. In the current design, the sphere was still quite narrow (basically only geography and inventions), however, none of the subjects had answered all the questions correctly. This finding suggests that there was no big expert in both inventions and geography, which made the CWM less powerful.

Second, an issue might be insufficient amount of robustness checks that were done due to technical limitations and time constraints. One concrete example is the performance analysis of the models with less people where only two sets of weighting questions were used.

Third, I did not find a way how to account for the dependence of the observations in the large sample and therefore, the results of the three applied tests had to be taken with caution. In further research, I suggest either to find a way how to account for this dependence or use a large sample size from the beginning so that no 'crowd-splitting' is necessary afterwards.

Subsequently, it is also important to supervise subjects during question collecting in order to make sure that if there is a difference in number of correct answers of online and offline respondents, it is due to better knowledge of people who used internet.

Finally, I believe that it would be useful to make sure that questions are not too difficult nor, on the other hand, too easy in further research. In the current design, apparently, this was the case, which lead to inconclusive results when studying the change in the number of people included in the models and the number of weighting questions.

Chapter 5

Conclusion

The crowd wisdom shows that mathematically combining opinions of people in a group delivers at least as good results as an individual answer of any member of that group. The concept has been developing since the beginning of the 20^{th} century and nowadays it finds use in many different spheres including prediction markets, prediction polls as well as crowd-sourcing.

There are many different ways how to aggregate group knowledge, some of them emphasize subjective confidence or past performance, some of them take into account, for example, peer evaluation. This thesis focuses on the performance comparison of three specific approaches: the simple average of answers, the surprisingly popular answer method (SPA) from Prelec et al. (2017) and the contribution weighted model (CWM) from Budescu and Chen (2014). The SPA determines the best answer as the one that is more popular than predicted and performs well when most of the people choose option they are familiar with but that is wrong. On the other hand, the CWM mathematically combines answers that are provided by people who make a positive relative contribution to the crowd performance. The main motivation for the algorithm comparison is as follows: Prelec et al. (2017) showed that the SPA outperforms the subjective confidence weights algorithm, especially in questions in which the majority of the crowd is wrong. As the CWM is also based on the subjective confidence and performs well for large sample sizes and various question topics, it is quite natural to check whether it can improve the confidence weights in Prelec's paper. Additionally, the simple average is included in the comparison in order to check that the models are properly constructed.

In order to obtain the data for the analysis, an anonymous quiz with 35 general knowledge questions was designed and distributed both in online and

in paper form to check for a different answering pattern. In total 102 responses were collected mostly from university students aged on average 23. As the budget for the research was limited, the subjects were motivated by a competitive element: each participant obtained, based on his or her answers, an objective accuracy score and a prediction score which could be afterwards compared with the scores of the other quiz participants on freely accessible website. After the data collection, the SA, SPA and CWM were constructed so that each of the algorithms gave answers to the 35 quiz questions. Since the CWM needed 10 questions to define the expertise, only 25 questions were used for the comparison of the models. Besides the three algorithms, the answers of people with a positive BTS score were also included in most of the comparisons in order to get more insights into the analysis.

To assess whether the number of correct answers delivered by the models was statistically different from chance I applied a binomial test. For the comparison of the algorithms, I used McNemar test, the matched pair sign test (only in some cases) and the t-test. The performance was further graphically evaluated also by F1 score and Matthews correlation coefficient. Besides that, the following topics were a subject of interest: the performance comparison of the CWM and the BTS with less subjects included in the models, relationship between the BTS score and the CWM score, analysis of cheating behavior in the online version of the quiz, whether there is a difference in performance if the sample size increases from 25 to 150 by splitting the large crowd into six smaller crowds and finally, how the number of weighting questions changes the performance of the CWM.

The results of the analysis showed that for any of the models, the high number of correct answers was not random. Additionally, for the small sample size there was no difference between the SA, SPA, CWM and the same held for the BTS comparison. The finding might be explained by several factors. First, I found a significant positive relationship between the quadratic and the logarithmic CWM score and the BTS score, which serves as a basis for the SPA and therefore, if the CWM performs well the SPA does so, too. Second, Fisher exact test proved that the higher number of correct answers in the online version of the quiz was non-random. Even though, more evidence would be needed to firmly conclude that cheating was present, the different answering pattern might have influenced the performance of the models so that the advantages of individual algorithms could not be sufficiently materialized. Finally, the sample size was too small to make any robust conclusions. For a larger sample with

$n = 150$, where the SA, SPA and CWM were compared by McNemar test, the matched pair sign test and the t-test, the number of correct answers for the SA was significantly lower than for the SPA and CWM when one version of 6 crowds and 20 different versions of weights were used. On the other hand, for one version of weights and 20 different 6 groups, the performance difference was significant and robust only for the SA and SPA comparison (SA again performed worse). Similarly as for small sample, there was no difference in the performance between the SPA and CWM in neither of the cases. To conclude, both the SPA and CWM perform equally but better than the SA for large sample size. However, the performance difference between the SPA and SA is more robust. Since the SPA has more stable results and does not need any specific number or domain of questions, it is in general better to use than the CWM that has more requirements for derivation. Regarding the supplementary analysis, I did not find any significant difference between the CWM and BTS performance when less subjects were used, similarly, the number of weighting questions did not significantly affect the outcome.

There are two main limitations of this study. The first one is insufficient incentivization of the subjects: with low motivation, only 35 questions could be used in the quiz which made the comparison of the models rather difficult. Therefore, using monetary rewards is highly encouraged in future research. The second issue was that the assumption about the independence of observations for McNemar test and the t-test was not met in the large sample size and therefore, the results of the tests had to be taken with caution.

Bibliography

Abaramowicz, M. and Henderson, M. T. (2006). Prediction markets for corporate governance. *Notre Dame L. Rev.*, 82:1343.

Arazy, O., Morgan, W., and Patterson, R. (2006). Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. *Proceedings of the 16th Workshop on Information Technologies and Systems. Milwaukee: WITS*, pages 79–84.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., et al. (2008). The promise of prediction markets. *Science*, 320(5878):877–878.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, 463(7279):294–295.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., and Mellers, B. (2016). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*.

Bickel, J. E. (2007). Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49–65.

Budescu, D. V. and Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., and Tetlock, P. E. (2016). Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses. *Decision Analysis*, 13(2):128–152.

Das, A., Gollapudi, S., Panigrahy, R., and Salek, M. (2013). Debiasing social wisdom. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–508. ACM.

Du, Q., Hong, H., Wang, G. A., Wang, P., and Fan, W. (2017). CrowdIQ: A New Opinion Aggregation Model. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Eickhoff, M. and Muntermann, J. (2016). Stock analysts vs. the crowd: Mutual prediction and the drivers of crowd wisdom. *Information & Management*, 53(7):835–845.

Forsythe, R., Nelson, F., Neumann, G. R., and Wright, J. (1992). Anatomy of an experimental political stock market. *The American Economic Review*, pages 1142–1161.

Gallaugher, J. (2012). Getting the most out of information systems. *Retrieved December*, 27:2014.

Gordon, K. (1924). Group Judgments in the Field of Lifted Weights. *Journal of Experimental Psychology*, 7(5):398.

Herzog, S. M. and Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2):231–237.

Hill, S. and Ready-Campbell, N. (2011). Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15(3):73–102.

Hosseini, M., Moore, J., Almaliki, M., Shahri, A., Phalp, K., and Ali, R. (2015). Wisdom of the crowd within enterprises: Practices and challenges. *Computer Networks*, 90:121–132.

Jain, R., Jain, P., and Jain, C. (2015). Behavioral biases in the decision making of individual investors. *IUP Journal of Management Research*, 14(3):7.

Jin, X., Gallagher, A., Cao, L., Luo, J., and Han, J. (2010). The wisdom of social multimedia: using flickr for prediction and forecast. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1235–1244. ACM.

King, A. J., Cheng, L., Starke, S. D., and Myatt, J. P. (2012). Is the true wisdom of the crowd to copy successful individuals? *Biology Letters*, 8(2):197–200.

Koriat, A. (2012). When are two heads better than one and why? *Science*, 336(6079):360–362.

Lehdonvirta, V. and Bright, J. (2015). Crowdsourcing for public policy and government. *Policy & Internet*, 7(3):263–267.

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

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276.

Parent, G. and Eskenazi, M. (2011). Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Twelfth Annual Conference of the International Speech Communication Association*.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *science*, 306(5695):462–466.

Prelec, D., Seung, H. S., and McCoy, J. (2013). Finding truth even if the crowd is wrong. Technical report, Tech. rep., Working paper, MIT.

Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535.

Prelec, D. and Seung, S. (2006). An algorithm that finds truth even if most people are wrong. *Unpublished manuscript*.

Simmons, J. P., Nelson, L. D., Galak, J., and Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15.

Sunstein, C. R. and Hastie, R. (2015). *Wiser: Getting beyond groupthink to make groups smarter*. Harvard Business Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Teevan, J. and Yu, L. (2017). Bringing the Wisdom of the Crowd to an Individual by Having the Individual Assume Different Roles.

Tetlock, P. E., Mellers, B. A., and Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355(6324):481–483.

Treynor, J. L. (1987). Market efficiency and the bean jar experiment. *Financial Analysts Journal*, 43(3):50–53.

Wagner, C. and Suh, A. (2014). The wisdom of crowds: impact of collective size and expertise transfer on collective performance. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 594–603. IEEE.

Wang, G., Kulkarni, S. R., Poor, H. V., and Osherson, D. N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, 8(2):128–144.

Wolfers, J. and Zitzewitz, E. (2004). Prediction markets. *The Journal of Economic Perspectives*, 18(2):107–126.

Wolfers, J. and Zitzewitz, E. (2006). Interpreting prediction market prices as probabilities. Technical report, National Bureau of Economic Research.

Appendix A

Appendix

Table A.1: Regressions of the quadratic CWM s. on the BTS s.: 1-5

VARIABLES	(1) CWMScore	(2) CWMScore	(3) CWMScore	(4) CWMScore	(5) CWMScore
BTSScore	0.198***	0.230***	0.220***	0.222***	0.221***
	(0.0506)	(0.0551)	(0.0508)	(0.0396)	(0.0648)
Gender_dummy	-0.00452	0.0228	-0.0223	0.0128	0.00648
	(0.0144)	(0.0175)	(0.0178)	(0.0146)	(0.0155)
Education_cat	0.00759	0.00443	0.0174	0.0148*	0.0154
	(0.00857)	(0.00949)	(0.0108)	(0.00805)	(0.00989)
Nationality_cat	0.00336	0.00376	0.00320	-0.00210	0.000667
	(0.00305)	(0.00292)	(0.00352)	(0.00271)	(0.00354)
Age	-0.00107	0.00246	0.00149	0.000611	-0.00134
	(0.00303)	(0.00304)	(0.00297)	(0.00271)	(0.00316)
Newshours	0.000662	-0.00115	-0.000696	-0.00155	-0.000680
	(0.00193)	(0.00206)	(0.00254)	(0.00163)	(0.00176)
Constant	-0.000706	-0.110	-0.0513	-0.0433	-0.00627
	(0.0690)	(0.0735)	(0.0663)	(0.0633)	(0.0696)
Observations	77	77	77	77	77
R-squared	0.203	0.281	0.301	0.415	0.216

Table A.2: Regressions of the quadratic CWM s. on the BTS s.: 6-10

VARIABLES	(6) CWMScore	(7) CWMScore	(8) CWMScore	(9) CWMScore	(10) CWMScore
BTSScore	0.257***	0.213***	0.283***	0.159***	0.0969**
	(0.0809)	(0.0427)	(0.0646)	(0.0527)	(0.0458)
Gender_dummy	0.0119	0.0410***	0.0162	0.0409*	0.0294*
	(0.0151)	(0.0149)	(0.0215)	(0.0208)	(0.0170)
Education_cat	0.0196*	0.000219	0.0194*	0.00815	0.00838
	(0.0102)	(0.00845)	(0.0106)	(0.0115)	(0.00934)
Nationality_cat	0.00175	0.00302	0.00356	0.00422	0.00419
	(0.00382)	(0.00268)	(0.00321)	(0.00455)	(0.00301)
Age	-0.00178	-0.000848	0.00335	0.00154	0.000403
	(0.00300)	(0.00259)	(0.00315)	(0.00399)	(0.00289)
Newshours	7.72e-05	6.47e-05	-0.000817	-7.55e-05	-0.000153
	(0.00171)	(0.00171)	(0.00331)	(0.00224)	(0.00160)
Constant	-0.0265	-0.0542	-0.156**	-0.135	-0.0931
	(0.0701)	(0.0637)	(0.0780)	(0.0908)	(0.0661)
Observations	77	77	77	77	77
R-squared	0.271	0.465	0.335	0.204	0.149

Table A.3: Regressions of the quadratic CWM s. on the BTS s.: 11-15

VARIABLES	(11) CWMScore	(12) CWMScore	(13) CWMScore	(14) CWMScore	(15) CWMScore
BTSScore	0.295***	0.0416	0.216***	0.110*	0.295***
	(0.0662)	(0.0586)	(0.0399)	(0.0567)	(0.0786)
Gender_dummy	0.0282*	0.0442**	0.0265	0.0212	0.0158
	(0.0158)	(0.0203)	(0.0206)	(0.0199)	(0.0160)
Education_cat	0.00559	0.00306	0.0135	0.0227*	0.0127
	(0.00922)	(0.0118)	(0.0114)	(0.0120)	(0.00988)
Nationality_cat	0.00767**	0.00428	0.00176	0.00113	0.00202
	(0.00363)	(0.00377)	(0.00343)	(0.00415)	(0.00407)
Age	0.00109	0.00364	-0.00192	-0.00263	-0.00773**
	(0.00317)	(0.00380)	(0.00387)	(0.00340)	(0.00330)
Newshours	0.000569	-0.00223	0.000735	-0.000172	9.41e-06
	(0.00200)	(0.00177)	(0.00217)	(0.00204)	(0.00178)
Constant	-0.118	-0.163*	-0.0301	-0.0303	0.121
	(0.0774)	(0.0855)	(0.0793)	(0.0811)	(0.0740)
Observations	77	77	77	77	77
R-squared	0.382	0.127	0.342	0.120	0.335

Table A.4: Regressions of the quadratic CWM s. on the BTS s.: 16-20

VARIABLES	(16) CWMScore	(17) CWMScore	(18) CWMScore	(19) CWMScore	(20) CWMScore
BTSScore	0.292***	0.228***	0.198***	0.109***	0.370***
	(0.0477)	(0.0365)	(0.0463)	(0.0289)	(0.0660)
Gender_dummy	0.0121	0.00891	0.0225	0.0267**	0.0226
	(0.0148)	(0.0141)	(0.0184)	(0.0132)	(0.0151)
Education_cat	0.0116	-0.00834	0.00406	0.0113	0.00536
	(0.00921)	(0.00789)	(0.0101)	(0.00824)	(0.00997)
Nationality_cat	0.000893	0.00386	0.00509	-0.000388	0.00448
	(0.00284)	(0.00275)	(0.00331)	(0.00248)	(0.00278)
Age	-0.000365	0.00459*	0.00402	0.00177	0.00310
	(0.00323)	(0.00244)	(0.00340)	(0.00326)	(0.00335)
Newshours	0.00118	0.00129	-0.000445	-9.62e-05	0.00140
	(0.00149)	(0.00191)	(0.00215)	(0.00125)	(0.00150)
Constant	-0.0467	-0.118**	-0.155*	-0.104	-0.146*
	(0.0728)	(0.0536)	(0.0779)	(0.0658)	(0.0770)
Observations	77	77	77	77	77
R-squared	0.361	0.395	0.287	0.328	0.511

Table A.5: Reg. of the logarithmic CWM s. on the BTS s.: 1-5

VARIABLES	(1) CWMScore	(2) CWMScore	(3) CWMScore	(4) CWMScore	(5) CWMScore
BTSScore	0.00470***	0.00497***	0.00493***	0.00479***	0.00515***
	(0.00104)	(0.00111)	(0.00115)	(0.000820)	(0.00138)
Gender_dummy	2.73e-05	0.000486	-0.000396	0.000284	0.000232
	(0.000307)	(0.000360)	(0.000384)	(0.000307)	(0.000335)
Education_cat	0.000179	0.000107	0.000373	0.000325*	0.000363*
	(0.000179)	(0.000194)	(0.000230)	(0.000171)	(0.000211)
Nationality_cat	7.69e-05	8.01e-05	6.78e-05	-4.43e-05	2.26e-05
	(6.44e-05)	(6.10e-05)	(7.58e-05)	(5.68e-05)	(7.51e-05)
Age	-2.15e-05	4.70e-05	4.68e-05	1.00e-05	-2.55e-05
	(6.23e-05)	(6.20e-05)	(6.38e-05)	(5.63e-05)	(6.58e-05)
Newshours	1.27e-05	-2.61e-05	-1.51e-05	-3.27e-05	-1.61e-05
	(4.21e-05)	(4.23e-05)	(5.55e-05)	(3.36e-05)	(3.85e-05)
Constant	-0.000294	-0.00225	-0.00157	-0.000900	-0.000456
	(0.00142)	(0.00149)	(0.00143)	(0.00132)	(0.00145)
Observations	77	77	77	77	77
R-squared	0.246	0.299	0.317	0.426	0.256

Table A.6: Reg. of the logarithmic CWM s. on the BTS s.: 6-10

VARIABLES	(6) CWMScore	(7) CWMScore	(8) CWMScore	(9) CWMScore	(10) CWMScore
BTSScore	0.00565***	0.00470***	0.00585***	0.00350***	0.00234**
	(0.00173)	(0.000918)	(0.00136)	(0.00117)	(0.00101)
Gender_dummy	0.000307	0.000935***	0.000345	0.000883*	0.000625*
	(0.000325)	(0.000311)	(0.000444)	(0.000472)	(0.000366)
Education_cat	0.000436**	2.56e-05	0.000396*	0.000229	0.000182
	(0.000216)	(0.000176)	(0.000217)	(0.000266)	(0.000203)
Nationality_cat	3.77e-05	6.83e-05	6.86e-05	7.39e-05	8.85e-05
	(8.11e-05)	(5.55e-05)	(6.67e-05)	(9.81e-05)	(6.48e-05)
Age	-3.24e-05	-2.71e-05	6.82e-05	2.27e-05	1.60e-05
	(6.38e-05)	(5.38e-05)	(6.52e-05)	(9.37e-05)	(6.33e-05)
Newshours	-2.21e-06	1.97e-06	-1.78e-05	-3.60e-06	-3.96e-06
	(3.73e-05)	(3.52e-05)	(6.86e-05)	(4.77e-05)	(3.48e-05)
Constant	-0.000797	-0.00110	-0.00318*	-0.00269	-0.00216
	(0.00149)	(0.00133)	(0.00161)	(0.00208)	(0.00144)
Observations	77	77	77	77	77
R-squared	0.285	0.497	0.334	0.190	0.164

Table A.7: Reg. of the logarithmic CWM s. on the BTS s.: 11-15

VARIABLES	(11) CWMScore	(12) CWMScore	(13) CWMScore	(14) CWMScore	(15) CWMScore
BTSScore	0.00656***	0.000982	0.00448***	0.00251*	0.00626***
	(0.00133)	(0.00136)	(0.000924)	(0.00135)	(0.00161)
Gender_dummy	0.000631*	0.00109**	0.000657	0.000547	0.000343
	(0.000324)	(0.000459)	(0.000464)	(0.000456)	(0.000332)
Education_cat	0.000126	0.000146	0.000337	0.000519*	0.000273
	(0.000189)	(0.000270)	(0.000265)	(0.000275)	(0.000205)
Nationality_cat	0.000161**	8.54e-05	3.47e-05	1.36e-05	4.61e-05
	(7.40e-05)	(8.48e-05)	(7.54e-05)	(9.44e-05)	(8.46e-05)
Age	1.70e-05	6.97e-05	-4.58e-05	-4.62e-05	-0.000157**
	(6.30e-05)	(8.94e-05)	(9.03e-05)	(8.28e-05)	(6.73e-05)
Newshours	9.54e-06	-4.96e-05	9.12e-06	-7.99e-06	-4.37e-08
	(3.99e-05)	(4.02e-05)	(4.68e-05)	(4.49e-05)	(3.78e-05)
Constant	-0.00241	-0.00365*	-0.000725	-0.00101	0.00236
	(0.00154)	(0.00196)	(0.00182)	(0.00194)	(0.00151)
Observations	77	77	77	77	77
R-squared	0.417	0.135	0.313	0.122	0.347

Table A.8: Reg. of the logarithmic CWM s. on the BTS s.: 15-20

VARIABLES	(16) CWMScore	(17) CWMScore	(18) CWMScore	(19) CWMScore	(20) CWMScore
BTSScore	0.00681***	0.00518***	0.00482***	0.00267***	0.00784***
	(0.000999)	(0.000789)	(0.00108)	(0.000736)	(0.00137)
Gender_dummy	0.000348	0.000299	0.000633	0.000679*	0.000485
	(0.000306)	(0.000307)	(0.000429)	(0.000342)	(0.000317)
Education_cat	0.000264	-0.000161	0.000164	0.000318	0.000126
	(0.000193)	(0.000170)	(0.000235)	(0.000211)	(0.000208)
Nationality_cat	3.35e-05	9.27e-05	9.81e-05	-1.47e-05	9.74e-05*
	(5.88e-05)	(5.73e-05)	(7.53e-05)	(6.50e-05)	(5.78e-05)
Age	-1.43e-06	9.97e-05*	7.68e-05	3.84e-05	6.67e-05
	(6.74e-05)	(5.18e-05)	(7.95e-05)	(8.48e-05)	(6.98e-05)
Newshours	2.10e-05	2.61e-05	-1.79e-05	-6.48e-06	2.69e-05
	(3.07e-05)	(4.16e-05)	(4.81e-05)	(3.09e-05)	(3.15e-05)
Constant	-0.00136	-0.00280**	-0.00339*	-0.00253	-0.00315*
	(0.00151)	(0.00113)	(0.00178)	(0.00171)	(0.00158)
Observations	77	77	77	77	77
R-squared	0.418	0.430	0.301	0.313	0.521