ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL
OF ECONOMICS

"Wisdom of the crowd: How to get rid of errors and biases at the aggregate level"

Master Thesis

By: Julia Heisig, 456610
Erasmus University Rotterdam, MSc Behavioural Economics

22 July 2018

Supervisor: Aurélien Baillon
Second Assessor: Benjamin Tereick

ABSTRACT

In many experiments it has been shown that the wisdom of the crowd can outperform individuals and sometimes even experts. In this thesis, I test two approaches to further improve the performance of four different mathematical aggregation models in forecasting football matches. Two unweighted models, the mean and median, and two weighted models, the Brier Weighted Model (BWM) and the Contribution Weighted Model (CWM). The BWM weighs forecasts of subjects based on individual expertise, while the CWM determines the weighting based on the contribution that a subject makes to the crowd's expertise. The first approach improves the models by modifying the BWM and CWM. This is done by decreasing the power of past performance for the first 10 events, which reduces the problem of requiring a track record without decreasing the prediction accuracy. The second approach increases the forecasting accuracy of the mean, median and BWM significantly by providing subjects with additional information on what three other subjects estimated. I show that this increase in accuracy is due to subjects having more doubts, which decreases both the bias of estimating too extreme values as well as fans estimating too optimistically for their team. This decrease in biases overweighs the anchoring bias due to the displayed estimates. In addition, a more thorough analysis of the CWM shows that different than stated in Budescu & Chen (2014) the model seems to only partly identify experts based on individual expertise.

Keywords: Wisdom of the crowd ◦ Forecasting ◦ Contribution Weighted Model ◦ Brier Weighted Model ◦ Debiasing

TABLE OF CONTENT

1 INTRODUCTION

When people make a decision, it is important for them to know what they can expect about the future. This holds for many fields such as politics, economics, sports or the weather. For this reason forecasting is a very relevant topic. Forecasts can either be made by individual experts or by a group of people. Asking only a few experts however can be a problem, because it is likely that they suffer from individual biases that affect the accuracy of their forecast (Mannes et al. 2014; Shanteau & Stewart 1992; Tetlock 2005). Therefore, it can be better to ask more people, to make use of the so called "wisdom of the crowd". There are various possibilities of how to ask more individuals, such as having a discussion forum with several experts, letting individuals vote on a decision, or aggregating the opinions of individuals mathematically (Chen et al. 2016). Due to problems of using discussion forums and voting rules such as the cost and complexity of organization, this thesis focuses on the mathematical aggregation of opinions.

The most well-known experiment of the wisdom of the crowd goes back to the statistician Francis Galton in 1907, who asked over 700 individuals to estimate the weight of an ox (Galton 1907-1; Galton 1907-2). As expected, most of the people were individually quite far away from the real weight. However, the median of all estimates was only 9 pounds away from the real weight, and the mean was even closer (1 pound distance to the true result). This surprising precision of the crowd's estimate is possible because although individuals might be biased and make errors, those biases and errors mostly cancel out at the aggregate level (Makridakis & Winkler 1983; Surowiecki 2004). That is why using the mean or median of estimates is a widely used instrument in forecasting (Hastie & Kameda 2005; Larrick & Soll 2006; Larrik, Mannes & Soll 2011; Soll & Larrick 2009). In addition to those simple unweighted models, there are also models that distribute different weights to the judges based on a track record or prior tests. One example for this is the Brier Weighted Model (BWM) which determines the weight based on the past performance by assigning Brier scores to each past estimate. In Budescu & Chen (2014), the authors come up with an alternative method: The Contribution Weighted Model (CWM). While also making use of past performance, the focus of the CWM is to identify those subjects who improve the crowd's forecast.

In Chen et al. (2016), the authors further improve the CWM as well as the BWM and mean model by additionally providing training in probabilistic reasoning and political analysis (the events concerned political events), and forming teams of 15 members who were able to state their opinions in an online platform. However, training and forming teams is costly, both from a monetary and also from a time perspective. That is why in my thesis, I want to extend the exploration of those models by investigating a cheaper way of improving the four aggregation models (CWM, BWM, median and mean). For that, two different approaches are used.

The first approach addresses a disadvantage of the BWM and CWM: In order to determine the weights, both need data about past performance. This means that when somebody wants to start using such a model for making a forecast, the models have a problem to already produce a stable forecast in the starting period. That is why for improving the performance of those models in my thesis the determination of the weights is modified. For this, I determine the weighting for the first 10 events not only based on the weighting that the BWM and CWM suggest, but combine it with a constant which consists of the unweighted mean. The ratio between those two components then linearly increases from 100% unweighted mean for the first event to 100% BWM/CWM from the tenth event onwards.

The second approach is an attempt to decrease systematic biases and increase expertise. Chen et al. (2016) showed that forming teams increased expertise. In addition, it led to subjects reorienting their mindset as well as concentrating on the analytical arguments for their estimation, which decreased systematic biases. As a result, they found that the prediction accuracy increased for all used models. I test if an increase in prediction accuracy can be achieved through a simpler method: Only displaying three estimates of random previous subjects. In addition, the drivers of the effect of this approach are investigated, such as if and where expertise increased and if biases decreased.

The structure of this thesis is as follows. Chapter 2 provides some conceptual background to the wisdom of the crowd and its principles as well as the four used aggregation models. In addition, it introduces possible cognitive and motivational biases during the estimation process, which are the tendency of individuals to estimate too extreme values, the optimism bias as well as the anchoring heuristics. Chapter 3 describes the dataset gathered by a questionnaire where 181 subjects estimated the outcomes of 33 football matches as well as

the general experimental design. In addition, chapter 3 also explains the two weighted models (BWM and CWM) in more detail and how the models compare to each other. Chapter 4 presents the results. First, it shows the advantage of the modification of the weighted models for the first 10 events: Being able to predict already from the very beginning, without having a lower prediction accuracy. Second, a comparison of the prediction accuracy of the models in the baseline and treatment condition shows that three out of four models perform significantly better in the treatment condition. The comparison also shows that different to Budescu & Chen (2014), the weighted models do not perform significantly better than the unweighted model. This could be due to the different area of events, but also due to a lack of power of this study. The comparison of the prediction accuracy demonstrates that the models appear to have different strengths: While the BWM is the most stable one, the median model leads to the highest maximum score, and the CWM to the highest median score across all events. In total, none of the models performs significantly different to the odds of a betting website. Next, a more detailed analysis of the effect of the treatment shows several findings. First, displaying three estimates makes all subjects (independent of the expertise) have higher mean scores. Second, people estimate less extreme values (0% and 100%), and more centric values (50%) when they have three estimates of previous subjects displayed. This might come from them having more doubts about their estimates. Third, the distance of estimates between fans of a team and the rest (non-fans) is lower when the subjects are in the treatment condition, however not significantly. Fourth, an analysis shows the existence of an anchoring effect of the mean of the three displayed values. Finally, the BWM and CWM are compared with each other in more detail to gain a better understanding on what they capture. The analysis shows first, that those subjects that receive a higher weighting from the CWM are not necessarily those with high individual expertise. Second, making too extreme forecasts appears to be less punished by the CWM than by the BWM. Afterwards, chapter 5 and 6 present the limitations of the experimental design, as well as a discussion of the findings and conclude.

2 CONCEPTUAL BACKGROUND

2.1 Wisdom of the crowd

Galton's example of guessing the weight of an ox is not the only one that shows the power of the wisdom of the crowd. In experiments where subjects had to estimate the number of coffee beans in a jar or the temperature in a room, the aggregated guesses came repeatedly surprisingly close to the true result (e.g. Galton 1907-1; Knight 1921; Lorge et al. 1958; Sunstein 2006; Treynor 1987). Sniezek & Henry (1989) argue that the wisdom of the crowd performs better than individuals (both inexperienced and experts), because it manages to capture the central tendency of a crowd that can be interpreted as knowledge/expertise. However, there are four conditions that must hold in order to have a wise crowd estimate (Larrik et al. 2011; Simmons et al. 2011; Surowiecki 2004).

First, the members of the crowd must have at least some expertise about the questions (Larrik et al. 2011). If the crowd does not have any knowledge, then the answers are just random, which does not lead to a combined estimate that is close to the true result. Yet, the literature findings are quite mixed on how much expertise is best for the accuracy of the result. On the one hand, Pachur & Biele (2007) show that experts perform better than inexperienced subjects. On the other hand, Andersson et al. (2005) show that both perform equally well and Gröschner & Raab (2006) even find that subjects with little expertise can outperform experts. The last is also known as the "wisdom of the ignorant crowd". Herzog & Hertwig (2011) made an experiment that explains why ignorant crowds can still be powerful. In the experiment, their subjects had to predict the outcomes of several tennis and soccer tournaments without having much knowledge. What they found was that the subjects instead relied on recognizing the teams or single players. They argue that recognition is not purely random, but instead reflects information that can be valuable for forecasting.

Second, the individuals in the crowd need to have diverse perspectives on the topic of the event, and also bring different expertise to the task (Larrik et al. 2011). If this holds, they will make different mistakes, which cancel out at the aggregate level. If the group is instead not diverse, the group is likely to make the same mistakes, which then remain at the aggregate level. The power of the lower errors can be seen at the following example. Let us assume that one asks some 15 year old girls who are friends for their guesses of the size of the Eiffel Tower. This is clearly not a diverse group. It can turn out well if they all guess

right. For example, if they are asked to estimate the size of the Eiffel Tower, and they have just been to Paris together, they all know the size. Hence, both their individual errors and the error of the crowd is equal to zero. However, if the crowd instead exists also of two more individuals, one who estimates the height of the Eiffel Tower twenty meters too low and the other one 10 meters to high, this changes. One can see that in this scenario the less diverse group performed better both on the aggregate and individual level. However, what one can also see is that for the more diverse group, the aggregate guess leads to a lower error than asking the individuals separately, since the errors of the two members partly cancel each other out (Surowiecki 2004). If the girl group mentioned earlier was instead asked a question that they all do not know, the more diverse group would hence improve the aggregate result. Page (2007) introduced the Diversity Prediction Theorem, which explains the above mentioned example. The theorem states that the collective error is equal to the average individual error minus prediction diversity. Therefore, the collective error is always equal or smaller than the average individual error. The effect of diversity based on personality types on accuracy in wisdom of the crowd models was also analyzed by Jain et al. (2011). In their study, they formed both very similar and very different couples. Then they asked them to estimate the number of M&M's in a jar, and also similar to this thesis if a football team wins a match. They found that for the football question the diverse teams were correct 42% of the cases, while the similar team was only correct in 32% of the cases. The same was found for the M&M estimation question, where the diverse couples were the closest to the true result. This experiment highlights the meaning of diverse opinions. Because of that importance, one study even tried to make each subjects think diverse themselves by providing them with contradicting scenarios (Herzog & Hertwig 2009).

Third, the individuals in the crowd must provide their answers independently (Larrik et al. 2011). If this is not the case, for example because individuals work in a group, they can be influenced to state a different answer than intended by feeling the group's pressure of choosing the same answer (Koehler & Beauregard 2006; Larrik et al. 2011). Also, individuals can be anchored by the answers of others, and afterwards not adjust sufficiently (Tversky & Kahneman 1974; Epley & Gilovich 2001). I will elaborate more on this in chapter 2.3.

Fourth and last, the judgements of all individuals of the crowd have to be aggregated to form a collective judgement. Lyon & Pacuit (2013) differentiate between three forms of aggregation: Group deliberation, prediction markets and mathematical aggregation.

When group deliberation is the chosen aggregation method, it means that the group members share their information with the other group members. There are two possibilities for information sharing: Unstructured and structured. In the unstructured variant, all subjects debate using the information that they have. However, this method is often criticized for even enhancing cognitive errors instead of cancelling them out, such as the tendency of individuals to ignore minorities (Lyon & Pacuit 2013). Therefore, it is especially known to be effective in brainstorming situations. The most famous example is an (unstructured) discussion in the polymath blog, which resulted in a new proof of the Hales-Jewett Theorem (Polymath 2012). The structured deliberation on the other hand tries to decrease the effect of such biases. One example for that is the Delphi method (Linstone & Turoff 1975). Here, individuals (mostly experts) answer questions in different rounds. In each round, an independent facilitator summarizes all forecasts of the previous round including its rationales anonymously. After a predefined stop rule (i.e. a specific number of rounds or stable results) the collective judgement is determined (Rowe & Wright 1999).

The second aggregation method that can be used are prediction markets. Rather than asking every subject for his opinion, a prediction market gives individuals the chance to trade contracts, whose payoff is connected to the event that is forecasted. For example instead of asking individuals what is the chance of Donald Trump getting re-elected as president, individuals have the chance to buy a contract. In the case that Trump is re-elected, the contract is worth 100€, and if not it is worth 0€. When transaction cost are ignored, this means that everyone should pay exactly that amount for the contract that reflects his subjective probability that the event is going to happen (if risk-neutrality is assumed). The clearing market price can then be interpreted as the aggregate probability of all investors (Manski 2006; Wolfers & Zitzewitz 2006). The prediction accuracy of those markets has been shown to be high (Arrow et al. 2008). However, although there are strong economic incentives in place, there are also biases such as the favorite longshot bias (Thaler & Ziemba 1988). This means that investors tend to overestimate events where the chance of occurring is close to 0 and on the other hand underestimate those events which have a very high chance (close to 1) of happening. In addition, it can be difficult to find enough investors for an event. And if there are too little investors, the market price can be vulnerable to manipulation (Lyon & Pacuit 2013).

Both group deliberation and prediction markets are very costly methods, since they require a high amount of organization. That is why the focus of this thesis is on the third aggregation method, the mathematical aggregation. The next chapter will explain this method more extensively and also explain the models that are used in this thesis.

2.2 Mathematical aggregation

Mathematical aggregation is probably the most commonly used aggregation method, and also the simplest one (Lyon & Percuit 2013). It means that a linear combination of the predictions of all crowd members is used to form one aggregate crowd judgement (Davis-Stober et al. 2014). This was proven to lead to high accuracy which outperforms the judgement of individuals (Davis-Stober et al. 2014; Galton 1907-1; Hahn & Tetlock 2006). In this thesis, I compare four mathematical aggregation models with regard to their prediction accuracy, (1) the mean and (2) median as well as the (3) BWM and (4) CWM.

Both the mean and the median model are using an unweighted aggregation of the predictions of the subjects. Armstrong (2001) recommends the mean as default model when there is only little information on the abilities of the subjects. However, using the average only makes sense if individuals are clustering around a central value. Lyon & Percuit (2013) show this using the following example of hypothetical estimates of the US GDP growth rates for the next decade:

a) -0.1%, 0.1%, 0.2%, -0-3%, 0.1%, 0.3%, 0.2%, -0.1%
b) -19.1%, 5.1%, 5.2%, 4.7% , -20.5%, 5.4%, 4.7%, 4.6%, 4.8%, 5.1%

For both examples, the average of all estimates is 0%. However, while in a) the estimates are also clustered around 0%, in b) it rather looks like they are clustered around 5%. This shows that the average is sensitive to outliers, in this case -19.1% and -20.5%. A better alternative here would be to use the median, which is the centric value with 50% of the estimates being lower and 50% of the values being higher. There are also other alternatives proposed, such as using the mode (most popular estimate) or the geometric mean, but in this thesis I will focus on the (arithmetic) mean and median.

Both the mean and the median have been criticized for often not leading to accurate forecasts, for example in case of systematic biases or when the crowd has only little expertise (Budescu

& Chen 2014; Simmons et al. 2011). If there is more information about the ability of the subjects, weighted models have been proposed as an alternative. Those models assign different weights to the predictions of the subjects based on criteria such as expertise (e.g. Aspinall 2010; Cooke 1991). This is based on the assumption that there is a high correlation of forecasts, such that subjects either mostly perform well or badly (Broomell & Budescu 2009). Cooke (1991) came up with an approach where he determines the weights of individuals based on pretests. However, I want to focus this thesis on approaches that are as easily applicable as possible and require only little cost. Hence, I decided to use two models that determine expertise based on the performance in past events. The Brier Weighted Model and the Contribution Weighted Model.

The Brier score was invented in 1950 to forecast two mutually exclusive binary events, where the probability of the two events has to add up to 1 (Brier 1950; Young 2010). They were first used for meteorology (e.g. rain or no rain), but later also for forecasting other events (e.g. Hvattum & Arntzen 2010; Peeters 2018). The Brier score signals the accuracy of a probability prediction by calculating the squared distance of the predicted probabilities to the true result. Hence, the lower the Brier score, the better the prediction. The Brier score can lie between 0 and 1.

The second weighted model used in this thesis is the Contribution Weighted Model which was developed by Budescu & Chen (2014). It compares the accuracy that the crowd would have with and without an individual. The more an individual improved the forecast (on average) in the past, the higher the weight that his probability estimates receive for future events. If an individual in contrast on average decreased the accuracy of the crowd's forecast, his estimates are excluded for future events. This is a difference compared to the Brier Weighted Model, where all estimates are used. In addition, the CWM rewards individuals who perform well when the majority of the crowd is wrong. This is particularly important if the judgement of different individuals turn out to be correlated, for instance because they rely on the same heuristics (Broomell & Budescu 2009). Normally, this makes it very difficult to forecast scenarios where the majority is wrong. However, giving those individuals a higher weighting that perform well in those scenarios makes the CWM more robust for such events than other models like the BWM. A more detailed description of the BWM and CWM follows in chapter 3.3.1. Since one goal of this thesis is to decrease the biases at the aggregate level, it is important to understand what biases exist when individuals make estimates.

2.3 Estimation biases

Since Tversky & Kahneman's seminal paper (1974), behavioural research has demonstrated that probability judgements of both laypeople and experts are subject to numerous biases. Recently, Montibeller & von Winterfeldt (2015) provided an overview with biases that distort forecasting results. Overall, they differentiate between cognitive and motivational biases.

First, cognitive biases are due to erroneous mental processes, leading to judgements that systematically violate normative rules (Gilovich et al. 2002; Kahneman et al. 1982; Montibeller & von Winterfeldt 2015). One example for this, which I also analyze in chapter 4.4.2, is the tendency of individuals to provide probability estimates that are too extreme as well as probability distributions that are too tight (Lichtenstein et al. 1971; Moore & Healy 2008). Dawes (1979) explains this effect psychologically with the desire of individuals to be able to predict certain events. He concludes that this desire translates into an implicit assumption that the event is actually easy to predict, which makes the individuals guess extreme values.[1]

Second, motivational biases distort decisions based on self-interest, social pressure or the organizational context (Montibeller & von Winterfeldt 2015; von Winterfeldt 1999). Motivational biases can both be conscious and unconscious. Since my thesis is about estimating the probability that a football team wins a match, I find one motivational bias particularly relevant: The wishful thinking/optimism bias. Being optimism-biased means the tendency of individuals to be overly optimistic about members of their group relative to others. This pattern can be expressed for example in ratings (Aronson et al. 2010; Taylor et al. 1981). Simmons (2011) found in a study where he asked different football fans to estimate the results of the matches that the fans of a team were significantly overestimating the chance of their favorite team winning. Hence, I expect this bias to be relevant for my thesis as well.

There is a lot of research experimenting with different debiasing techniques to reduce or in the best case eliminate biases such as the ones mentioned above. In my thesis, I use the displaying of three estimates of previous subjects as a debiasing technique. This a very simplified version of the one used by Chen et al. (2016), where the subjects were grouped into teams and could see the estimates of their team members. They argue that this leads to

[1] This bias is also often connected with the overconfidence bias, which is however not covered in this thesis.

subjects reorienting their mindset and concentrating on the analytical arguments for their estimation. If the subjects concentrate on the analytical arguments, they are less prone to biases. However, using this debiasing technique might lead to the introduction of a new bias: The anchoring effect. Discovered by Tversky & Kahneman (1974), the anchoring effect means that individuals tend to unconsciously use provided information as a reference point and adjust too little from it. For example, they made a study where one group of subjects were asked to compute the result of 1*2*3*4*5*6*7*8 in 5 seconds, and the other group was asked to compute the same, but in reverse order: 8*7*6*5*4*3*2*1. Since 5 seconds is not enough to compute the correct result (without a calculator), the subjects had to guess the result. Most individuals did so after starting with the first multiplications, which was then used as an anchor. Tversky and Kahneman (1974) found that the median estimate of those subjects who had the increasing sequence was 512, while the median guess of the ones who had the calculation in a decreasing order was with 2250 significantly higher.

In order to keep this bias as low as possible, Montibeller & von Winterfeldt (2015) suggest two different debiasing techniques. First, they propose the provision of not only one but multiple anchors. This is also in line with Block & Harper (1991), who showed that only providing individuals with one number leads to a stronger anchoring effect than providing them with multiple numbers. In addition, Montibeller & von Winterfeldt (2015) suggest to not choose one anchor that is constant over all subjects, but instead provide everyone with a different anchor. The application of both is discussed in the next chapter.

3 METHOD

3.1 Experimental design

For gathering data, I ran a survey asking subjects to estimate the chance of a national football team winning a friendly match. There was a baseline and treatment condition in the experiment. In the baseline condition, the subjects did not receive any additional information. In the treatment condition, the subjects were provided with three random estimates of previous subjects. This is in line with the debiasing techniques discussed in 2.3, since multiple numbers are used, which are in addition randomly chosen for every subject. The subjects were randomly distributed to one of the two conditions, so that the design of the experiment was between subjects.[2] It is important that the subjects are randomly assigned to either treatment or baseline group, because otherwise there would be a high chance of having selection bias. For instance, very confident individuals might not want to know the estimates from others while less confident individuals might be more open to having more information (Angrist & Pischke 2009).

As there are many estimates required for determining the CWM and BWM weights, it is important to choose a topic where subjects do not mind answering many questions. Also, as stated in 2.1, a crowd can only be wise if the subjects in the crowd have some knowledge. Hence, the topic needs to be one where most individuals have at least some knowledge of. In addition, it is necessary that the different questions are expertise-transferable (Simmons et al. 2011), so that the questions should come from the same domain. Otherwise, the model determines expertise using events from one field which might not be relevant for the domain of the event that needs to be forecasted. Also, it can be difficult to keep the biases of the different estimation questions constant over all questions. This would be a problem for the CWM and BWM methods, since they rely on the similarity of questions. If the questions are not correlated, the determination of experts from previous questions might not be of use for the other forecasting tasks. Hence, I decided to choose football as a topic, because it fulfills all three criteria. First, almost half of the people globally are interested in football, and even 20% play football themselves (Nielsen Sports 2018). Second, most people have some knowledge of football. To also ensure that the third criteria of transferable expertise holds, it

[2] However, since not all subjects completed the study, there are 83 subjects in the baseline condition and 88 subjects in the treatment condition.

is important to use similar matches that are also taking place around the same time. Thus, I decided to use international friendly matches of national football teams. With the number of matches, there is a trade-off between gathering as much data as possible and keeping the survey short enough to ensure that subjects complete it. In their first study, Budescu & Chen (2014) used the data of subjects who answered at least 10 questions, although they mentioned as a limitation that this number is too little. That is why I decided to ask for all national football matches from May 26th to June 2nd, which are 33 questions (BBC 2018). Since one match was cancelled, there are 32 matches left for the data analysis.

In Budescu & Chen (2014), the authors used two different approaches. First, they asked their subjects to estimate the chance for an event happening vs. not happening. An example for this is: "The average mortgage rate for a 30-year fixed-rate loan in the US will be above 4.5% before 30 March 2012.". Second, they asked subjects to be more specific and estimate the probabilities of different categories. For example they did not only ask the subjects what is the chance that the mortgage rate is above 4.5%, but what is the probability that it is between 4.0 and 4.5% and so on. There were 7 to 22 categories per question (Budescu & Chen 2014; ECB 2018).[3] Since I want to keep it as simple as possible and thus also easy to understand for the subjects, I use the first approach.[4] Thus, for their estimate, the subjects can choose a probability estimate between 0 to 100. The typical question for the baseline group looks like this:

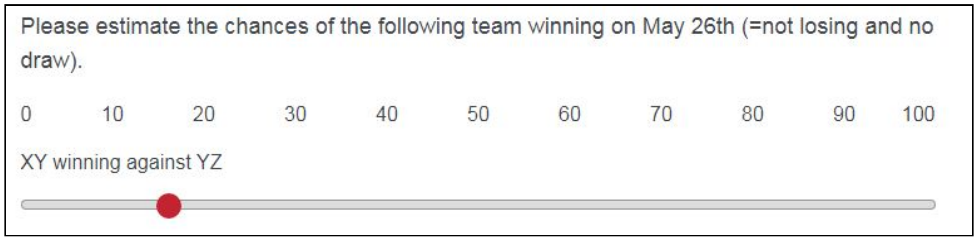


Fig. 1: Exemplary estimation question for the baseline group

For the treatment group, exactly the same information is given, however there is a second sentence added that provides them with information about three estimates of other random subjects:

[3] In Figure A.1 in Appendix A there are two screenshots of exemplary events of the two studies. For their second study, they did not gather data themselves but instead used a dataset from the ECB.

[4] In contrast to Budescu & Chen (2014) I only asked the subjects to provide the probability estimate for the event to occur and not for the probability estimate that it does not occur. I did that to make the survey shorter, since the additional probability estimate does not add additional value/information.

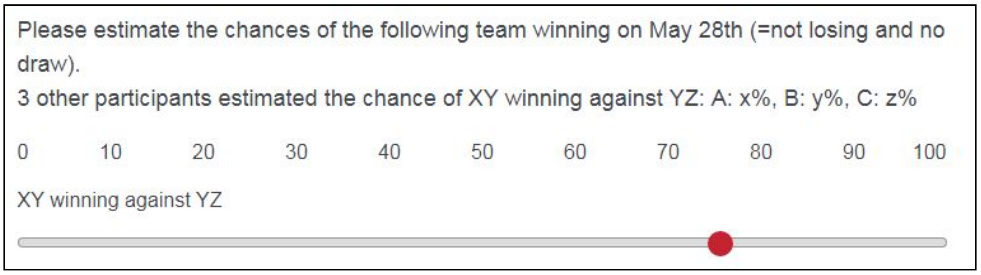


Fig. 2: Exemplary estimation question for the treatment group

I decided to not display the questions in a random order, because this is also not the case when subjects are asked questions over time. Of course this might mean that they spend more energy on the first questions, but I would also expect this effect when subjects are repeatedly asked to state their estimates. Also, this provides the opportunity that for calculating the BWM and CWM weights for an event, always the same previous events are used. In addition, it might provide more information to know that a national team just played two days ago against another team, which is the case for several teams. The question is displayed in a slider format, because this format was found to be easier to use on devices with a touchscreen and also seems more like a game (Cape 2009). In addition, Chen et al. (2016) also use this format style for their survey.

In the optimal scenario, I would have been able to program a survey in such a way that it automatically randomly selects three of the answers of previous subjects. However, this was technically not possible for me in the short time period. Instead, I first ran the survey with 10 subjects. Afterwards, I used the data I received here to generate fifty random combinations of those 10 subjects, from all 720 (10*9*8) possible combinations, excluding the possibilities of having the estimate of a subject more than once. When a subject was assigned to the treatment condition, the subject was randomly assigned to one of the fifty possible treatment variants. I ensured that all of those fifty possible variants are displayed evenly. In Chen et al. (2016), the authors did not change the groups of subjects in one "season". Hence, I decided to also always display the estimates of the same three subjects across all 33 questions. The subjects were informed about that (see Tab. A.4 in Appendix A). This enabled the subjects to have learning effects from the information over time, for example to pay more attention to only one or two estimates if one had obvious wrong estimates in past questions.

Because I only used the estimates of 10 subjects, I tried to make sure that those subjects are as diverse as possible. Hence, they consist of 6 different favorite teams (often congruent to nationality), 50% males and 50% females. Also, the age of those subjects ranges from 19 to 53 years, with a mean of 32 years. 4 of them are students, 3 work part-time and 2 work full-time and 1 is unemployed.

3.2 Incentives and subject pool

Asking subjects to estimate 33 football matches is time-intense and can get monotone because the type of question does not vary. Hence, there needs to be an incentive that motivates subjects first to participate, but also to put effort into the probability estimation process. It would be too expensive to pay everyone who fills out the survey, so I introduced a point system based on the Brier score, which is the standard scoring rule in most studies (Peeters 2018). For every question, the subjects could score between 0 and 10.000 points, depending on the squared distance from the true event. To ensure that they understand the rule and its quadratic form, I provided them with the following explanation and calculation example, which can also be seen in Tab. A.3 and Tab. A.4 in Appendix A.

"For every estimate, you can earn up to 10.000 points. The points are given based on how close you are to the true result. If the team wins, the real result is 100, and if the result is draw or losing, the true result is 0.
(More information to the rule I'm using: (100 points - distance from true result)².
Example: If you estimate that the chance of Germany winning against Japan is 80% and Germany wins, you get 6400 points (100 - 20 distance)²=80². If Germany does not win (either lose or draw), you get 400 points (100 points-80 points distance)².) "

To determine the winner of the experiment, I added up the points for each subject for all of the estimated matches after the result of the last match was public. Then, I translated the points into lottery tickets, such that the better a subject performs, the higher the chance to win the prize of 30€. Afterwards, I randomly chose one lottery ticket as the winning one, and determined the owner of that ticket as the winner. In other studies, incentives are often given for the result of only one random question (Bardsley et al. 2009). However that would give

the incentive for subjects to estimate very extremely (either 0 or 100) for every question, in order to be closest to the true result, which is not desired. Hence, I chose to instead use a lottery, where the chance to win is higher for those subjects who perform well, but the ones who do not perform well still have a chance to win. Also, the chance of winning is increasing linearly with their performance, which should provide them with an incentive to exert effort in the task (Baltussen et. al 2012; Bardsley et al. 2009).

Since the stake is relatively low with only a chance of winning 30€, in combination with a relatively monotone task, I aimed for a survey duration of about 15 min. There are 33 estimation questions, where I would think that subjects need 20-30s per question. In addition, there are three demographic questions, which I would estimate to take 10s to answer each, a short introduction as well as the explanation of the rules. Summing all up (having 1 min each for the introduction and explanation), the survey was planned to take 16.5min. However, the median duration of the experiment was 9.3 minutes, which can be explained by learning effects of the subjects. In the limitation section, I will elaborate about other possible reasons.

I aimed to collect 160 responses (80 per condition), based on Wagner & Suh (2014), who introduced the rule of thumb for wisdom of the crowd studies of having around 50-150 subjects, dependent on the amount of factors in the study. In the experiment, 225 subjects in total participated, over a span of three weeks.[5] However, only 181 subjects also finished the tasks. After excluding the first 10 subjects that are used to determine the three displayed estimates, there are 171 subjects remaining for the data analysis. Besides estimating the chance of a national football team winning, the subjects were also asked to fill out some demographic questions. They were asked for their age, gender and employment status, in order to be able to (superficially) analyze how diverse the subject pool is. The youngest of the 171 subjects is 17 years old, and the oldest 75 years, with a mean of 31 years. On average, the age of the subjects varies from the mean by 13 years. Overall, 60% of the subjects are male and 40% are female. The two largest employment groups with 43% each are student and full-time employed. 6% are employed part-time, 3% are retired, 1% is unemployed and 2% have a different employment status. The majority (61%) of the subjects say that they sometimes watch national football matches, 26% of the subjects say that they watch every football match possible, and 13% never watch national football matches. The subjects are

[5] One could argue that the ones who answered the survey closer to the match had more information available and were thus able to predict the matches better. However, first there is only little prior press attention for international friendly matches and second even the last day of survey collection was 2.5 weeks away from the first actual match.

fans from 23 different national teams. The country with the most fans in the experiment is Germany (59%), followed by the Netherlands (17%).

In order to have enough subjects, I asked 75 students of my Applied Statistics tutorial to participate (around 50 participated). In addition, I shared the survey on several student facebook groups, with prior work colleagues and family and friends. I mainly distributed it in Germany and the Netherlands, but also tried to reach as many subjects as possible in other countries. Also, I tried to reach as many different age groups as possible. By doing this, I wanted to achieve a high diversity of the subjects. That is also the reason why I decided to have an online and not offline survey, to be able to gather the estimates from subjects from different locations. The survey was constructed with the online survey platform Qualtrics.

3.3 Analysis

3.3.1 Brier Weighted Model and Contribution Weighted Model

After I collected the data as described above, I used four different aggregation models to make one collective judgement for each event. Since the BWM and CWM are more complex, I am going to explain them in detail.

The Brier score for a subject j ($j = 1, ..., J$) for an event i ($i = 1, ..., I$) is a quadratic loss function of the following form (Brier 1950):

$$BS_{ji} = (o_i - p_{ji})^2$$

with o_i being the observed true result for each event i with $o_i = 1$ if the event occurs and $o_i = 0$ otherwise, and p_{ji} being the predicted probability for the outcome for an event i by a subject j. The Brier score can be between 0 and 1, with 0 being the highest score and 1 being the lowest score possible.

In Budescu & Chen (2014), the authors then rescale the Brier score BS_{ji} to a score S_{ji} that is also used later to compare the scores of all models.

$$S_{ji} = a + b\, BS_{ji}$$

Here, they set $a = 100$ and $b = -100$ in order to have a score between 0 and 100 with 0 being the lowest score possible and 100 being the highest score possible. Afterwards, the average score $\overline{S_{jk}}$ up to the last event $i = k$ is calculated.

$$\overline{S_{jk}} = \frac{1}{k} \sum_{i=1}^{k} S_{ji}$$

In order to forecast the next event $k+1$, the forecasts of all subjects are weighted based on their average score $\overline{S_{jk}}$ until the last event k. Hence, the weighting $q_{BWM,jk+1}$ is based on the absolute performance of subjects in all k past events. The higher the average score for the past k events, the higher the weighting $q_{BWM,jk+1}$ of the individual subject j.

$$q_{BWM,jk+1} = \frac{\overline{S_{jk}}}{\sum_{j=1}^{J} \overline{S_{jk}}}$$

The Contribution Weighted Model also determines the weighting for an upcoming event based on the past performance, which is measured by the same scoring rule as the BWM. However, the scores are not determined for each subject, but instead for the whole crowd:

$$S_i = a + b\,(o_i - \overline{p_i})^2$$

with $\overline{p_i}$ being the predicted probability for an event to occur using the average of the estimates of all subjects. Afterwards, the contribution C_{ji} of a subject ($j = 1, ..., J$) for event i is determined as the difference between the crowd's score S_i and the crowd's score S_i^{-j} without subject j 's estimated forecast $\overline{p_i^{-j}}$.

$$C_{ji} = S_i - S_i^{-j} \text{ with } S_i^{-j} = a + b\,(o_i - \overline{p_i^{-j}})^2$$

The contribution can either be positive, negative or equal to zero, which means that the crowd either has a better, worse or exactly the same performance for that event with the subject than without him. Then, the subject's total contribution up to the last event $i = k$ ($k = 1, ..., I$) is calculated similar to the BWM, using the average of all contributions C_{ji} up to the last event $i = k$.

$$\overline{C_{jk}} = \frac{1}{k} \sum_{i=1}^{k} C_{ji}$$

To determine the weighting of opinions of the subjects for a future event $k+1$, the average contribution of a subject $\overline{C_{jk}}$ of all previous events up to the last event k is used. Only the forecasts of those subjects are taken into account who have a higher mean contribution than the threshold ε. In the paper from Budescu & Chen (2014), the authors choose $\varepsilon = 0$ for their main analysis, so they exclude all negative contributors with $\overline{C_{jk}} < 0$ from the forecast of that event. However, other values of ε are also possible and used by the authors for additional analyses. The forecast is then determined based on a weighted average of all subjects with $\overline{C_{jk}} > \varepsilon$. Same as Budescu & Chen (2014), I also use $\varepsilon = 0$, so that all subjects with a positive contribution $\overline{C_{jk}} > 0$ are included in my analysis. The weighting $q_{CWM,jk+1}$ of a forecast p_{jk+1} of an individual j for upcoming event $k+1$ is proportionally increasing with his mean contribution to past events $\overline{C_{jk}}$.

$$q_{CWM,jk+1} = \frac{max(\overline{C_{jk}}, 0)}{\sum_{j=1}^{J} max(\overline{C_{jk}}, 0)}$$

Both the Brier weighted and Contribution Weighted Model base the weighting of forecasts on the past performance of subjects. Hence, they both require a track record of past events. The smaller the track record, the higher the chance that the good performance in past events was due to pure luck instead of expertise. To take that into account, I used the proposed idea of Budescu & Chen (2014) to modify the starting weights. For that, the weights w_{jk+1} of the different subjects are calculated by using a weighted average based on previous performance q_{jk+1} (determined by the BWM or the CWM) and a constant $V = \frac{1}{J}$ representing an equal weight of all subjects.

$$w_{ji} = \omega_i q_{jk+1} + (1 - \omega_i)\frac{1}{J}$$

The weight ω_i of those two components evolves linearly over time/events starting with a weight $\omega_i = 0$ for the first event, since there is no track record available, and finally $\omega = 1$ for the 10th event, from which point on the forecast only relies on past performance. However, as a robustness check, I do the same until the 15th and 20th event.

The above described method of determining weights for the BWM and CWM differs from the two methods that are used in Budescu & Chen (2014). The first one proposed by them is a jackknifing procedure, where all events except the one that is forecasted are used to determine the weighting. However, I think that this procedure is not very useful for practical purposes, since here only past events can be used, while the results of upcoming events are still unclear. In the second proposed method the new events are included in a dynamic fashion, similar to my approach. However, this method is only applied in combination with a track record of 104 events. This is not possible with my dataset. Instead, with the modification of decreasing the power of past performance for the first 10 events I want to ensure that also those events with little track record can already be forecasted without relying too much on luck.

3.3.2 Comparison of the models

For comparing the different models, I use the rescaled Brier score that I explained in 3.3.1 and that was also used in Budescu & Chen (2014) and Chen et al. (2016). After determining the score that each model received for each event, I compare the minimum, maximum, mean and median scores across all events, as well as the standard deviation and compare those values across the models.

To ensure comparability to Budescu & Chen (2014) and Chen et al. (2016), for comparing the improvement of prediction accuracy I am using the proportional relative improvement (PRI) which they used as well. It is defined as:

$$PRI = \frac{score\ treatment - score\ baseline}{(100 - score\ baseline\ model)}$$

If there is no improvement in score, the PRI is equal to 0, if there is a decline in score, the PRI is negative, and for a perfect improvement (to the highest score possible), the PRI is equal to 1. The reason I am using this measure instead of a simple relative improvement is that it takes into account that the room for improvement from for example 99 to 99.1 is different than from 15 to 15.1 because the scale naturally ends at a score of 100.

3.4 Hypothesis development

Based on the theoretical background and the explained method, this chapter establishes some hypotheses for the later analysis. I start with the effect of modifying the BWM and CWM by making the weighting of the forecasts of different subjects for the first 10 events to a lesser degree dependent on past performance. That is because the fewer events the models base their weighting on, the more likely it is that it is based on luck instead of expertise. However, disregarding information from past performance can also decrease the prediction accuracy. In total, I expect those two factors to cancel out, so that the prediction accuracy remains unchanged, but becomes more stable (less variance).

H1: The prediction accuracy of the modified BWM/CWM does not differ from the unmodified BWM/CWM, while the variance is lower for the modified versions.

Next, I expect all models to perform either similarly or better when having three estimates of previous subjects displayed. This would also be in line with the findings of Chen et al. (2016), who found that grouping subjects increased the scores that signal prediction accuracy. I expect this finding for two reasons.

First, seeing three estimates from other subjects increases the information available, which increases their expertise. I expect this to show in overall higher individual scores. I expect this effect to be the highest for subjects with little expertise who I expect to benefit the most from the additional information. With respect to the four models, higher expertise of the subjects is valuable for all of them. However, especially the mean model benefits from the improvement of subjects with little expertise, since this one is sensitive to outliers.

Second, I expect the treatment to reduce the optimism bias as well as the tendency to estimate extreme values. This is because subjects see diverse estimates from other subjects that might differ from their 'normal' answer. Hence, I expect them to question their answers more and be more analytical and less intuitive. This is also in line with the explanation of the findings of Chen et al. (2016). However, I also think that displaying three estimates of previous subjects can increase the anchoring bias, which has a negative effect on the prediction accuracy. In total, I expect the debiasing effect to be higher than the increase of the anchoring heuristics. Similar to above, I expect both subjects with little and high expertise to benefit from this, but especially those with little expertise, where there is more room for

improvement with the 'normal' answer. I expect all models to benefit from a decrease in biases.

H2: The prediction accuracy of all models is either unchanged or higher when subjects see three estimates from previous subjects.

As a next analysis, I compare the prediction accuracy of the different models to see which model leads to the most precise forecast. For the accuracy of the forecast, I compare the mean, median as well as the minimum and maximum of the scores of the models across all events, analogously to Budescu & Chen (2014) and Chen et al. (2016), who tested the prediction accuracy of the CWM against the BWM and mean.
I expect a similar ranking as in those two studies, which would mean that the CWM performs best. That is because it does the best job identifying those subjects in the crowd that increase the crowd's expertise, and it benefits from having a group with more expertise. When the forecasted event is highly correlated with the past events that are used to determine the weighting, this advantage is low, because then it is likely that the herd is correct. However, when the event is not highly correlated with the past events, the CWM performs better than the other models. That is because this method determines the subjects that increase the crowd's expertise depending on those events where the majority was wrong and their opinion deviated from the herd (Budescu & Chen 2014). I expect the BWM to perform second best. This expectation is different to the result of Chen et al. (2016), where the mean model performed better. However, I think that the modification of decreasing the power of the past performance for the first events will improve the stability of this model and improve it to perform better than the mean and median. Overall, I expect the two weighted models to perform best because by using data from past events they hedge against uninformed forecasts. I think that the CWM does that even better since it excludes those subjects who perform badly, while the BWM includes all subjects.

H3: I expect the two weighted models to perform better than the two unweighted models.
H3a: The CWM has the highest prediction accuracy, followed by the BWM.

Last, I compare the BWM and CWM in more detail, in order to understand what exactly the CWM captures. In Budescu & Chen (2014) the authors state that the CWM identifies the "experts" of the crowd. However, this seems contradicting because those subjects with a high

individual expertise are identified by the BWM. Since the CWM performed in both studies of Budescu & Chen (2014) as well as in Chen et al. (2016) better than the BWM, it seems to capture something different than individual expertise, which is the contribution to the crowd's expertise. Hence, I expect that those individuals that receive a high weighting in the CWM are not the same that receive a high weighting (above average) in the BWM.

H4: The subjects who receive a high weighting in the CWM are not the same who receive a high weighting in the BWM.

4 RESULTS

4.1 Effect of the modification of the BWM and CWM

I start my analysis by analyzing the impact of the modification of the two weighted models BWM and CWM. I do this by adding a constant consisting of the unweighted mean, so that the weighting for the first 10 events is not only determined by the BWM and CWM, but instead by a combination of both. This decreases the influence of past performance on the weighting for the events where there is only a small track record available. Here, I analyze how the scores change, if they change significantly and also if they get more stable. As a robustness check I did the same for the first 15 and for the first 20 events. The results are very similar and can be found in Table B.1 in Appendix B.1.

Tab. 1 Effect of modification of BWM/CWM on scores of models
The table compares the scores of the BWM and CWM in the modified and unmodified version. This is across the first 9 events, where the weighting in the modified version is not fully determined by the BWM or CWM yet. For the comparison of the standard deviation (SD), the first event is excluded, to make the modified and unmodified versions comparable.

Model	Modified	E1	E2	E3	E4	E5	E6	E7	E8	E9	SD
BWM	No	0	91.8	87.2	63.9	97.0	98.6	53.4	21.9	89.9	27.0
	Yes	89.5↑	91.3↓	87.0↓	64.1↑	96.9↓	98.6→	53.5↑	21.9→	89.9→	26.9↓
CWM	No	0	94.8	85.8	57.9	98.0	99.5	49.1	17.3	88.5	29.3
	Yes	89.5↑	92.0↓	85.9↑	63.0↑	97.2↓	99.0↓	51.2↑	19.0↑	88.6 ↑	28.3↓

Table 1 shows the difference in scores for the first 9 events between the modified and unmodified BWM and CWM. One can see several advantages of the modification. First, it is now possible to already make a forecast for the very first event. Second, in both the BWM and CWM, the three lowest scores are either higher or at same level with the modification. However, the three highest scores are also lower or at the same level. Hence, it appears as if the modified models are more stable. In total, it looks as if there is no difference between the scores, which is also confirmed by a Wilcoxon test. The test shows that the null-hypothesis that the scores of the two models do not differ significantly cannot be rejected at a 5% significance level (p_{BWM}=0.58; p_{CWM}=0.40).[6] Hence, there is support for the first part of

[6] For this, I run a paired-sample Wilcoxon test (paired by event), with the following hypothesis:
H0: μ_unmodified=μ_modified H1: μ_unmodified<μ_modified.
I used a non-parametric test because there are not enough observations to assume a normal distribution of the scores. In addition, the null-hypothesis of normal distribution was rejected at a 5% significance level for all models except one using a Shapiro-Wilk test. Because of consistency reasons, I used a non-parametric test.

Hypothesis 1, that stated that the difference in scores is not significantly different. In addition, one can see that the variance is lower for the modified versions than for the unmodified versions, which indicates a higher stability. However, the difference is not significant at a 5% significance level (p_BWM=0.50, p_CWM=0.46).[7] Hence, there is no support for the second part of Hypothesis 1, which stated that the variance is lower for the modified models than for the unmodified models. However, because of the advantage of having one additional event to forecast without having a lower performance, I use the modified version for both models for the further analysis.

[7] The hypothesis for the F-test is: H0: σ^2_modified=σ^2_ unmodified H1:σ^2_modified<σ^2_ unmodified.

4.2 Effect of displaying three estimates on prediction accuracy of models

As a second analysis, I test if the scores of the four aggregation models are significantly higher when using the estimates from the subjects who in addition see three estimates from previous subjects.

Tab. 2 Effect of treatment on prediction accuracy of models
The table compares the minimum, maximum, mean and median score as well as the standard deviation of the four models across the 32 events. It compares the scores of the models between the baseline and treatment condition and also depicts the proportional relative improvement from the baseline to the treatment condition (see chapter 3.3.2). The highest minimum, maximum, mean and median score as well as the lowest standard deviation across all models and conditions are highlighted in bold.

Model	Condition	Min	Max	Mean	Median	SD	p[8]
Mean model	Baseline	20.6	98.6	76.5	81.7	17.4	
	Treatment	23.3 ↑	98.5 ↓	77.3 ↑	81.4 ↓	17.2 ↓	0.03**
	PRI	+3.3%	-10.9%	+3.4%	-1.6%		
Median model	Baseline	15.4	99.2	75.5	79.3	19.4	
	Treatment	14.4 ↓	**99.4** ↑	77.3 ↑	80.9↑	19.4 →	<0.01**
	PRI	-1.2%	+25.0%	+7.3%	+8.4%		
BWM modified	Baseline	26.8	97.0	77.0	81.6	15.3	
	Treatment	**29.5**↑	97.7↑	**77.7**↑	81.4↓	**15.1**↓	0.03**
	PRI	+3.7%	+23.3%	+3.0%	-1.1%		
CWM modified	Baseline	23.7	97.2	76.0	79.2	16.5	
	Treatment	27.9↑	97.8↑	77.0↑	**82.5**↑	16.5→	0.20
	PRI	+5.8%	+21.4%	+4.2%	+15.9%		

* = p<0.1, ** = p<0.05 *** = p<0.01; n=32

Comparing the two conditions baseline and treatment shows that using the estimates of the treatment group leads to significantly higher scores than using estimates from the baseline group for all models except for the CWM, where the change is not significant at a 5% significance level. In total, there is support for Hypothesis 2 of either unchanged or higher scores. A possible explanation why the treatment did not have a significant positive effect on the CWM is that this model already does the best job identifying the subjects that

[8] For all of the aggregation models, I am testing the following hypothesis: H0: μ_baseline=μ_treatment H1: μ_baseline<μ_treatment by using a paired-sample Wilcoxon test, comparing the scores paired by event. I used a nonparametric test since the normality assumption of normal distribution was rejected at a 5% significance level using a Shapiro-Wilk-test.This is likely to be due to a right-skewness of the data.

compensate for biases. This means that there is less room for improvement. Additional explanations due to limitations of the experimental setting will be elaborated in chapter 5.1.

Next I compare the prediction accuracy of the different models. Different than expected, the BWM of the treatment condition performs best when comparing the mean score. Besides that, it also has the highest minimum score which in combination with the lowest standard deviation makes this model appear the most stable. When comparing the median score, the CWM performs best. Last, the median model has the highest maximum score. This seems intuitive, since this model is the least sensitive to outliers, so that it can also lead to extreme values if the majority of the crowd estimates so. This is also reflected in the lowest minimum score of the four models. For a more detailed analysis, the distribution of the scores of the models is shown in Fig. B.2.1 in Appendix B.2. In addition, in Fig. B.2.2 and Fig. 2.2.3 in Appendix B.2 show the development of the scores over the events. Hypothesis 3a stated that it is expected that the CWM performs best. In total, six Wilcoxon tests paired by events show that the null-hypothesis that the CWM scores are not different to the scores of the other models cannot be rejected at a 5% significance level.[9] [10] Hence, as expected from Table 2, there is no support for Hypothesis 2a. However, since this model suffers the most from the low number of events, the high median score could be a sign that with more events it might perform better than the other models.

Finally, I compare the performance of the two weighted models (mean and median) with those of the two unweighted models. While the two weighted models have lower minimum scores, the two unweighted models have higher maximum scores, both independent of the condition. For the mean and the median scores, neither the weighted nor the unweighted models are clearly better, but the unweighted BWM leads to the overall highest mean score and the CWM to the overall highest median score. Lastly, the two unweighted models have the lowest standard deviation, which could be a sign of a higher stability of the models. To see if the weighted models have significantly higher scores, I run several Wilcoxon tests, comparing one by one each weighted model with an unweighted model for

[9] For the six (three per condition) paired tests, I am testing the following hypothesis: H0: μ_cwm=μ_other H1: μ_cwm<μ_other by using a paired-sample Wilcoxon test, comparing the differences paired by event. I used a nonparametric test since the normality assumption of normal distribution was rejected at a 5% significance level using a Shapiro-Wilk-test.This is likely to be due to a right-skewness of the data. I use three paired Wilcoxon tests, since at a Mann Whitney U test information about the events would get lost.

[10] Results: p_baseline_cwm_mean=0.62, p_baseline_cwm_mean=0.84, p_baseline_cwm_bwm= 0.86, p_treatment_cwm_mean=0.62 , p_treatment_cwm_mean=0.76, p_treatment_cwm_bwm= 0.81.

both the baseline and treatment condition.[11] [12] However, the null-hypothesis that there is no difference in the scores of the weighted and unweighted models cannot be rejected at a 5% significance level for any of the tests. Hence, there is no support for Hypothesis 3 that stated that it is expected that the weighted models perform better than the unweighted models. A reason for this might be that 32 events are still too little events for the two weighted models, so that they would outperform the other models at a later stage.

4.3 Comparison of the prediction accuracy of the models to betting odds

To put the scores of the different aggregation methods into a better context, I compare it with the scores of a betting website. There is empirical evidence that betting odds are the most accurate forecast that is publicly available for many sports, such as football (e.g. Forrest et al. 2005; Song et al. 2007). Hence, I want to test if the betting odds of the website Oddsportal.com lead to significantly different scores than the four models.[13] For that, I checked the betting odds on the website during the data gathering phase, in order to have the same time distance to the event for both the subjects and the betting company. Because the betting websites use odds instead of probabilities and there is also a bookmaker's margin included in the betting odds, I first transformed them into probabilities. For that, I used the basic normalization approach (Strumbelj 2013). As a first step, the inverse odds $\gamma = (\gamma_1, \gamma_2, ..., \gamma_n)$ are calculated using the odds $\sigma = (\sigma_1, \sigma_2, ..., \sigma_n)$ with $n = 3$ possible outcomes (team 1 winning, tie, team 2 winning) with the condition of $\sigma_i \geq 1$ for all events $i = (1, ...I)$.

$$\gamma_i = \frac{1}{\sigma_i}$$

As a second step, the booksum β is determined, with $\beta - 1$ being the bookmaker's margin.

$$\beta = \sum_{i=1}^{n} \gamma_i$$

[11] The hypothesis for this test is: H0: μ_weighted (BWM or CWM)=μ_unweighted (mean or median model) H1: μ_weighted (BWM or CWM)>μ_unweighted (mean or median model) by using several Wilcoxon tests paired by event. I used a nonparametric test since the normality assumption of normal distribution was rejected at a 5% significance level using a Shapiro-Wilk-test. I used several Wilcoxon tests, since at a Mann-Whitney-U test information about the event would get lost.

[12] Results: p_bwm_median_baseline= 0.06, p_bwm_mean_baseline=0.89, p_cwm_median_baseline= 0.24, p_cwm_mean_baseline=0.64, p_bwm_median_treatment= 0.54, p_bwm_mean_treatment=0.73, p_cwm_median_treatment= 0.58, p_cwm_mean_treatment=0.62.

[13] I chose this website because it was the one who predicted the most football events that were used in the study. Since the matches were friendly matches, other major betting website providers often did not offer bets on the matches, or if only on few of the 33 selected ones.

Third and last, the outcome probability p_i can be determined, with the set of values p_i summing up to 1.

$p_i = \frac{\gamma_i}{\beta}$

Unfortunately, for many matches there were no betting odds available, so that I can only compare 14 events. I am using the modified version of the models, since otherwise there would only be 13 events to compare the scores with. Also I am using the treatment condition, since those scores were better than the ones from the baseline condition.

Tab. 3 Performance of models compared to a betting website
The table compares the minimum, maximum, mean and median score as well as the standard deviation of four selected models with those of a betting website across 14 events. The four models consist of the modified version of the BWM and CWM in the treatment condition, as well as the mean and median model of the treatment condition. The weights for the modified BWM and CWM were determined not only based on those 14 events, but instead of all 32 events. The highest minimum, maximum, mean and median score as well as the lowest standard deviation across all models and conditions are highlighted in bold.

Model	Min	Max	Mean	Median	SD
Betting website	**32.4**	92.8	74.6	77.7	**16.7**
Mean T	23.3 ↓	98.5 ↑	**74.8** ↑	**84.0** ↑	24.5 ↑
PRI	-13.5%	+79.1%	-3.2%	+14.9%	
Median T	14.4 ↓	**99.4** ↑	73.9 ↓	**84.0** ↑	28.0 ↑
PRI	-25.1%	+93.3%	-6.3%	+13.7%	
BWM modified T	29.5 ↓	97.6↑	74.1 ↓	78.0 ↑	20.4↑
PRI	-14.2%	+79.7%	-3.5%	+15.1%	
CWM modified T	27.9 ↓	97.8 ↑	74.0 ↓	77.9 ↑	22.1 ↑
PRI	-27.4%	+95.0%	-12.8%	-0.1%	

The betting website seems to make more stable predictions, since the minimum score is higher as well as the standard deviation lower than for all other models. This also leads to a higher average, since there are only 14 events so that one minimum event already has quite a large impact on the average. However, the median score is higher for all four models than for the betting website. In addition, I test with a Wilcoxon test paired by event if the scores of the

four models differ from those of the betting website.[14] In total, the null-hypothesis that the scores do not differ significantly cannot be rejected at a 5% significance level.[15]

In addition, it seems surprising that the mean and median model perform so well with regard to the median score. This comes from both models, but especially the median model, having overall more extreme probability estimates for the events, which can also seen at the lower minimum and higher maximum scores. Since for those 14 events there were more events where the models could achieve high scores, the median scores are also high. However, it is questionable if this trend would also hold for more events.

[14] The hypothesis is: H0: μ_model=μ_betting website H1: μ_model≠μ_betting website.
A non-parametric test was used since the number of observations is very low so that a normal distribution of the data cannot be assumed. This is reinforced by a Shapiro-Wilk test, where the null-hypothesis of a normal distribution can be rejected at a 5% significance level.

[15] p_mean=0.78, p_median=0.73, p_bwm=0.73, p_cwm=0.83.

4.4 Understanding the effect of displaying three estimates

4.4.1 Effect of displaying three estimates on individual scores

In 4.2 one could see that the overall prediction accuracy of three out of four models was significantly higher when three estimates of previous subjects were displayed. However, it did not become clear how exactly the forecasting performance of the subjects improved. To analyze this, I first want to compare all individual scores S_{ji} between the two conditions.

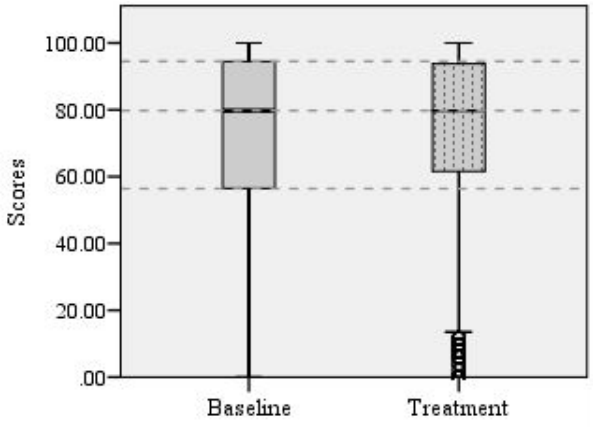


Fig. 3 Boxplots of all individual scores S_{ji}
The boxplot shows the distribution of all individual scores S_{ji} across all 32 events. It compares those subjects who were in the baseline condition with those in the treatment condition.

While the median individual scores remained unchanged between the two conditions as can be seen in Figure 3, the mean is 2.6% higher for the treatment condition. However, especially the 25% quartile of the treatment condition is remarkably higher (9.1%) than the baseline condition. This goes in hand with a smaller interquartile range, which shows that there are fewer low scores across all events. Hence, the scores are more stable on a high level for the treatment condition than for the baseline condition. There are two potential explanations for this finding. First, it might be that there are relatively fewer low-performing subjects, while there is no difference between the conditions for the high-performing subjects. This would mean that only some subjects improve. Second, it could be that the low scores become less on an individual level, so that subjects in the treatment condition have higher mean scores than in the baseline condition. This would mean that everyone improves. To understand better if only some or everyone improves, I run another analysis comparing the individual mean scores $\overline{S_{ji}}$ between the two conditions.

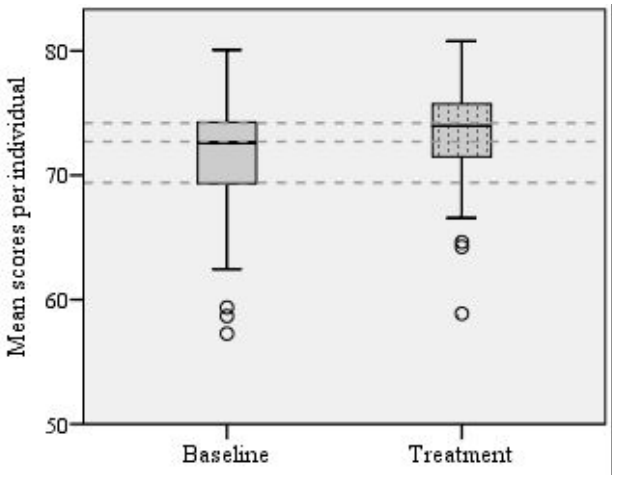


Fig. 4 Boxplots of the individual mean scores $\overline{S_{ji}}$
The boxplot shows the distribution of the mean individual scores $\overline{S_{ji}}$ per condition (baseline and treatment) . The mean individual scores $\overline{S_{ji}}$ are calculated based on the individual scores S_{ji} for the 32 events.

Figure 4 shows that the 25%, 50% and 75% quartile of the mean scores per subject are all higher in the treatment condition than in the baseline condition. In total, the individual mean scores are significantly higher for the treatment condition than for the baseline condition (p=0.003).[16] This suggests that the effect of the treatment cannot be explained with fewer weak-performing subjects. Instead, the scores for the treatment condition are higher for all levels of expertise. This is also supported by a more detailed analysis on the group level that can be found in Table B.3 in Appendix B.3.

As a next step, I want to understand if the systematic biases of estimating too extreme values and the in-group optimism bias are present in my study and if they decrease with the treatment.

4.4.2 Effect of displaying three estimates on estimating extreme values

First, I start with the bias of providing too extreme probability estimates. For that, I analyze the calibration curves of the subjects. When an individual is well-calibrated, then of all those events that he estimated to happen with a chance of for example 80%, 80% will actually occur in the long run. The calibration curve of such a perfectly calibrated individual

[16] The hypothesis is: H0: μ_baseline=μ_treatment H1: μ_baseline<μ_treatment.
A non-parametric Mann Whitney U test was used since the number of observations is very low so that a normal distribution of the data cannot be assumed. This is reinforced by a Shapiro-Wilk test, where the null-hypothesis of a normal distribution can be rejected at a 1% significance level. A Mann-Whitney U test was used since there is no pairing possible because of the between-subject design.

is represented by the identity line, and it is shown in Figure 5.[17] With the help of calibration curves, I want to understand if the subjects in my study are providing too extreme estimates. If this is the case, it would show in a significantly higher distance to the identity line for extreme values than for non-extreme values. To further understand if there is a difference between the baseline and treatment group, I make two calibration curves, one for each condition.

To determine the calibration curves of the subjects in my experiment, I group the estimates of all subjects in decimals, e.g. all estimates between 0% and 10% are in one group. This is the same procedure as used by Lichtenstein, Fischhoff & Phillips (1971). Afterwards, for all 10 groups the average of the estimated probability is then compared to the actual proportion that the events occurred. Fig. 5 below shows the calibration curves for the treatment and baseline condition.[18]

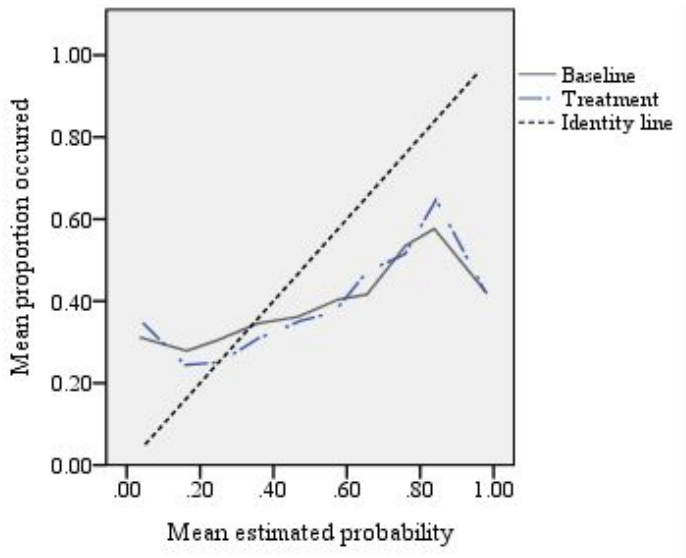


Fig. 5 Calibration curves vs. identity line
The figure depicts the relationship between the mean estimated probability of a group of probability estimates (e.g. 0%-10%) for an event to occur, and the mean proportion that the event actually occurred. This relationship is represented by the calibration curves, which are shown for the subjects of both the baseline and treatment condition. In addition, an identity line is shown.

The graph confirms the expectation of having a higher difference to the identity line for the extreme estimates 0% and 10% and 90% and 100% than for the other, non-extreme values. For example from all events that were estimated to happen with a chance of 4.5% (average of the group 0.0-0.1), more than 30% actually occurred. This is also shown by the

[17] The identity line is defined by f(x)=x.
[18] In Table B.4 in Appendix B.4 a similar analysis is done comparing the positive and negative contributors of the CWM.

kink after 0% and before 100%. A t-test shows in addition, that for both the baseline and the treatment condition the distance from the identity line is significantly higher for extreme values than for non-extreme values ($p_{baseline} = 0.003$, $p_{treatment} = 0.007$).[19] This provides evidence for the existence of the bias to provide too extreme probability estimates for both conditions.

Next, I want to see if this bias is lower for the treatment condition. Therefore, I compare the distance of the calibration curves to the identity line. While for the estimate of 100%, the calibration curve of the treatment condition is closer to the identity line than the calibration curve of the baseline condition, it is further away for the estimate of 0%. Hence, there is no clear tendency that the bias of providing too extreme estimates decreased.

However, besides being better-calibrated at estimating extreme values, the treatment might also have an effect of lowering the bias by making less subjects estimate those extreme values in general. This would not show in Figure 5. Hence, as a next step I look at the frequencies of all estimates.

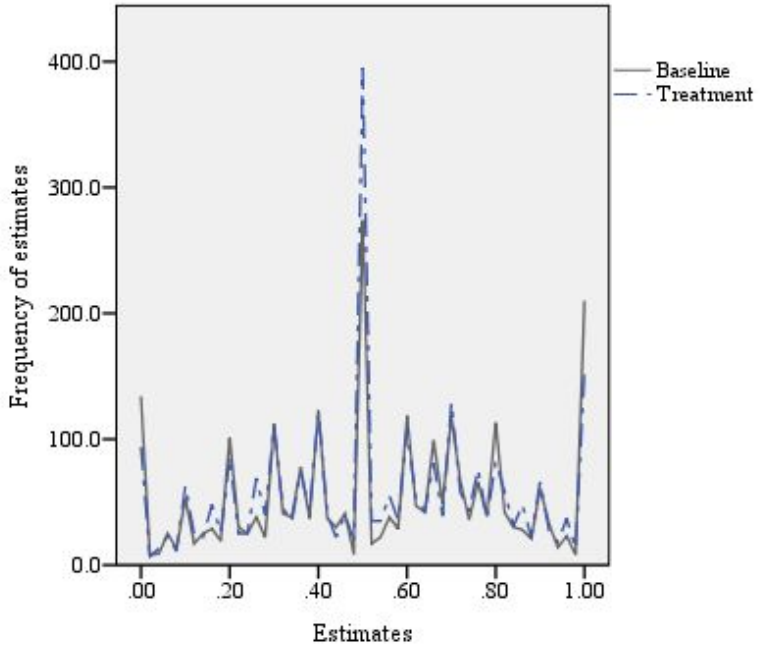


Fig. 6 Histogram of estimated probabilities
The graph shows the absolute frequencies of all probability estimates for the subjects of the baseline and treatment condition.

In the histogram above, one can see a difference between the estimates of the baseline and treatment group. The frequency for estimating 0% and 100% is notably lower for

[19] The null-hypothesis of a normal distribution could not be rejected at a 5% significance level using the Shapiro-Wilk test. The hypothesis tested in the two-sample t-test with equal variances is: H0: μ_distance_extreme=μ_distance_non-extreme, H1: μ_distance_extreme>μ_distance_non-extreme.

subjects in the treatment condition than for those in the baseline condition, so they seem to be less certain to estimate extreme. This higher doubt also shows in the higher frequency of estimating 50% of the subjects in the treatment condition than in the baseline condition. The table below compares the relative frequencies of those three values between the two conditions.

Tab. 4 Relative frequency of estimating extreme or centric values across treatment
The table shows the relative frequency that the the probability estimates 0%, 50% and 100% were predicted, as well as the difference between the baseline and treatment group.

Condition	0%	50%	100%
Baseline	4.9%	9.7%	7.5%
Treatment	3.2%↓	13.1%↑	5.2%↓
(Difference)	-34.7%	+35.1%	-30.7%

Table 4 confirms this impression and shows the relative differences of those frequencies. Hence, one can conclude that the treatment of displaying three estimates makes subjects estimate less extreme and more centric values. Hence, it seems as if the bias to estimate too extreme values is lower for subjects of the treatment group than those from the baseline group. The argument of doubt is reinforced by the time that the subjects of the two conditions take to complete the survey. While the median duration for subjects of the baseline condition is 8.5 minutes, the median duration for subjects of the treatment condition is with almost 10 minutes significantly higher at a 5% significance level ($p=0.03$). However, the subjects of the treatment condition also have to read more text.

For even deeper insights on which groups in particular have a different estimation behavior between the two conditions, Table 5 presents an analysis of the relative frequencies across the different groups of subjects.[20]

[20] In Appendix B.5 there is an additional analysis showing the relative frequencies of probability estimates that are multiples of 10, comparing different group of subjects.

Tab. 5 Relative frequency of estimating extreme or centric values across groups
The table shows the relative frequency that the probability estimates 0%, 50% and 100% were predicted by each group. It shows for each group the relative frequency of the probability estimates divided into baseline and treatment condition. For each group (e.g. Top 10 BWM) the counterpart of this group (e.g. Others (Not Top 10) BWM) is shown as a next group.

Groups	0%	50%	100%
Top 10 BWM[21] Baseline	1.5%	12.8%	2.1%
Top 10 BWM Treatment	2.1%	23.0%	2.7%
Others (not Top 10) BWM Baseline	5.4%	20.3%	8.3%
Others (not Top 10) BWM Treatment	3.4%	25.3%	5.5%
Top 10 CWM[22] Baseline	7.0%	6.5%	7.9%
Top 10 CWM Treatment	3.0%	19.5%	7.0%
Others (not Top 10) CWM Baseline	4.9%	19.4%	7.5%
Others (not Top 10) CWM Treatment	3.3%	25.8%	5.0%
Positive contributors[23] (CWM) Baseline	5.8%	16.0%	8.5%
Positive contributors (CWM) Treatment	3.3%	29.4%	3.5%
Negative contributors (CWM) Baseline	4.1%	22.5%	6.6%
Negative contributors (CWM) Treatment	3.2%	20.9%	6.6%

Interestingly, the trend is not the same across all groups. While the subjects of most groups have relatively less extreme estimates (0% and 100%) in the treatment condition than in the baseline condition, the opposite is observable for the Top 10 BWM. When comparing the relative frequency of the estimate 50%, there is an increase from the baseline to the treatment condition observable for all groups. However, this increase is especially notable for the Top 10 BWM, Top 10 CWM and positive contributors, so for all subjects who either possess high individual expertise and or improve the crowd's expertise. Those results reinforce the argument that the treatment leads to all subjects doubting their answer, independent of expertise. This is also related to my next analysis, where I want to see if the

[21] This group consists of the 10 subjects of a condition (baseline or treatment) that received the 10 highest weightings at the BWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) BWM".
[22] This group consists of the 10 subjects of a condition (baseline or treatment) that received the 10 highest weightings at the CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) CWM".
[23] This group consists of those subjects of a condition (baseline or treatment) that received a positive weighting at the CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition who were excluded from the weighting of the CWM are part of the group "Negative contributors CWM".

treatment makes fans of a team doubt their answer, and instead be less optimistic about their team.

4.4.3 Effect of displaying three estimates on the optimism bias

First, I want to test if I can see a trend for the optimism bias, so if the estimated probability for a team winning is significantly higher for subjects who are fans of that team than those who are not fans of that team.[24] Since there are only enough fans to be able to run a test from Germany (59% of all subjects) and the Netherlands (17%), I am going to test only for those two national teams.

Tab. 6 Optimism bias
The table shows the average probability estimates for two matches as well as the standard deviation (SD) of those estimates, both divided by fans and non-fans.

Match	Fan	n	Mean	SD	p[25]
Austria winning against Germany	Fan Germany	101	13.91	18.04	
	No Fan Germany	70	17.3	19.19	0.099*
Slovakia winning against the Netherlands	Fan Netherlands	29	17.24	18.47	
	No Fan Netherlands	142	28.39	21.36	0.01**

* = p<0.1, ** = p<0.05 *** = p<0.01

As it can be seen in the table above, the fans of Germany estimated on average a chance of 13.91% that Austria is winning against Germany, while the non-fans estimated this chance to be 17.3%. When testing for a statistical difference, the null-hypothesis that the estimate made by fans of Germany is equal to the estimate made by non-fans can be rejected at the 10% significance level. Similar to the fans of Germany, the fans of the Dutch national team estimate the chance of Slovakia winning against the Netherlands lower than non-fans (17.24% vs. 28.39%). In contrast to the fans of the German team, the difference is significant even at the 1% significance level. However, it is still difficult to conclude that this is due to optimism bias. It might also be that the German/Dutch fans are better informed and thus estimate the chance of the German/Dutch team better. Only comparing it with the actual result is not helpful here, since this does not represent the "true probability" from only one

[24] The subjects were asked as a part of the demographic questions, to select their favorite team, see Table A.2 in Appendix A.

[25] The null hypothesis of a normal distribution was rejected at 1% significance level using a Shapiro-Wilk test, so I decided to use a non-parametric test. For that I used an Independent Samples Mann-Whitney U test, with the hypothesis: H0: μ_fan=μ_nofan H1: μ_fan≠μ_nofan. I used that because there was no pairing possible.

event. If there were more matches of one team, it would thus be better to compare in addition if the scores of fans are significantly lower than those of non-fans.

Next, I want to see if the difference of estimates between fan and non-fan is reduced by the treatment of displaying three estimates of other subjects. Table 7 below presents an overview of the results as well as the statistical tests.

Tab. 7 Effect of treatment on optimism bias
The table shows the average probability estimates for two matches as well as the standard deviation (SD) of those estimates, both divided by fans and non-fans. It differentiates between subjects of the baseline and treatment condition.

Match	Condition	n	Fan	Mean	p[26]
Austria winning against Germany	Baseline	46	Fan Germany	12.0	
		37	No Fan Germany	17.5	
	Treatment	55	Fan Germany	15.5 ↑	0.33
		33	No Fan Germany	17.0 ↓	0.83
Slovakia winning against the Netherlands	Baseline	19	Fan Netherlands	16.6	
		64	No Fan Netherlands	28.4	
	Treatment	10	Fan Netherlands	18.5 ↑	0.36
		78	No Fan Netherlands	28.4 →	0.85

* = p<0.1, ** = p<0.05 *** = p<0.01

When comparing the fans of the baseline and treatment condition with each other and also the non-fans, it gets clear that the treatment condition had a larger effect for the fans. For the Netherlands, the mean estimate of the fans increased by 8.8% from 17.0 to 18.5, while the mean estimate for those subjects who were not fans of the Dutch team did not change at all. A similar trend can be observed for the fans of the German team: While the mean estimate of fans increases by 29.2% from 12.0 to 15.5, the mean estimate of those subjects who were not fans of the German team only decreased by 2.9% from 17.5 to 17.0. However, none of the differences is significant at a 5% significance level. Still, the general trend shows a reduction in the distance of estimates between fans and non-fans. Hence, this is a promising avenue for future research done with more subjects and especially more matches of the same team.

[26] The null hypothesis of a normal distribution was rejected at 1% significance level for both matches (using a Shapiro-Wilk test), so I decided to use a non-parametric test. Because there was no pairing possible because of the between subject design. I use a Mann-Whitney U test. The hypotheses that are tested are: H0: μ_fanbaseline=μ_fantreatment, H1: μ_fanbaseline<μ_fantreatment H0: μ_nofanbaseline=μ_nofantreatment, H1: μ_nofanbaseline≠μ_nofantreatment.

4.4.4 Anchoring effect of displaying three estimates

As shown in the previous chapters, the treatment has the effect of debiasing and increasing the prediction accuracy of aggregation methods. However, it can also create new biases, such as the anchoring bias. As mentioned before, I tried to decrease this effect by displaying not only one but three estimates of other subjects. However, in Whyte & Sebenius (1997), it was shown that there was still an anchoring effect although several numbers were used. Instead, the subjects were anchored by the mean of the presented anchors. This was explained by subjects perceiving the mean of the given numbers as consensus of this group. Although there is no group in my experiment, I am also testing for the existence of an anchoring effect of the average of the three displayed estimates. Of course, there are more possibilities for anchoring effects, such as the median, the minimum or maximum, the middle number, etc. of the three estimates. However, because figuring out which one leads to the highest anchoring effect is not the focus of this thesis, I only test for a possible anchoring effect of the mean of the three displayed estimates. Since the subjects in the treatment group (n=88) got 50 different versions of three estimates presented, I am using an approach similar to the one used by Ariely et al. (2003), who also had different anchors for all subjects. They separated the group in two parts: Below and above the median anchor. Then they tested, if the group with the above-median anchors also has significantly higher estimates than the group with the below-median anchors. I am doing the same, and separate the 88 subjects in the treatment condition in two groups. However, in contrast to Ariely et al. (2003), I define the two groups new for each of the 32 events, since the estimates displayed also vary with each event. For example for the first event, the median anchor (with the anchor being the average of the three presented estimates) is 66.7. Hence, the two groups here are first those subjects who got three estimates displayed with a mean of below 66.7 and second those subjects who got three estimates displayed with a mean of above 66.7. For the second event, the groups are newly formed based on if the average of the three estimates that were displayed is below or above 35.8, and so on.

A paired-sample t-test shows that the mean estimates of subjects that see a high anchor estimate a significantly ($p<0.001$) higher probability for the team winning than those

subjects who have a low anchor.[27] [28] The difference in mean estimates can also be seen in the graph below.

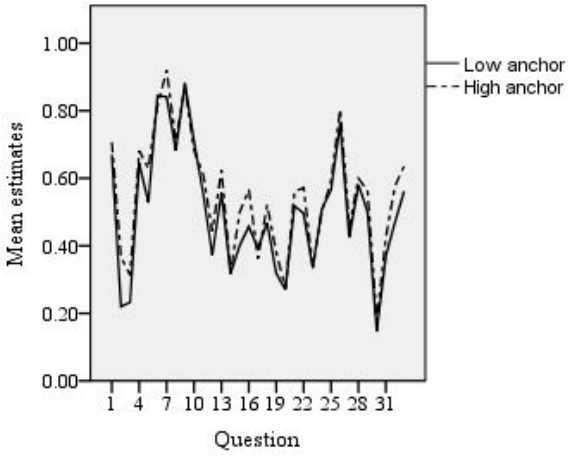


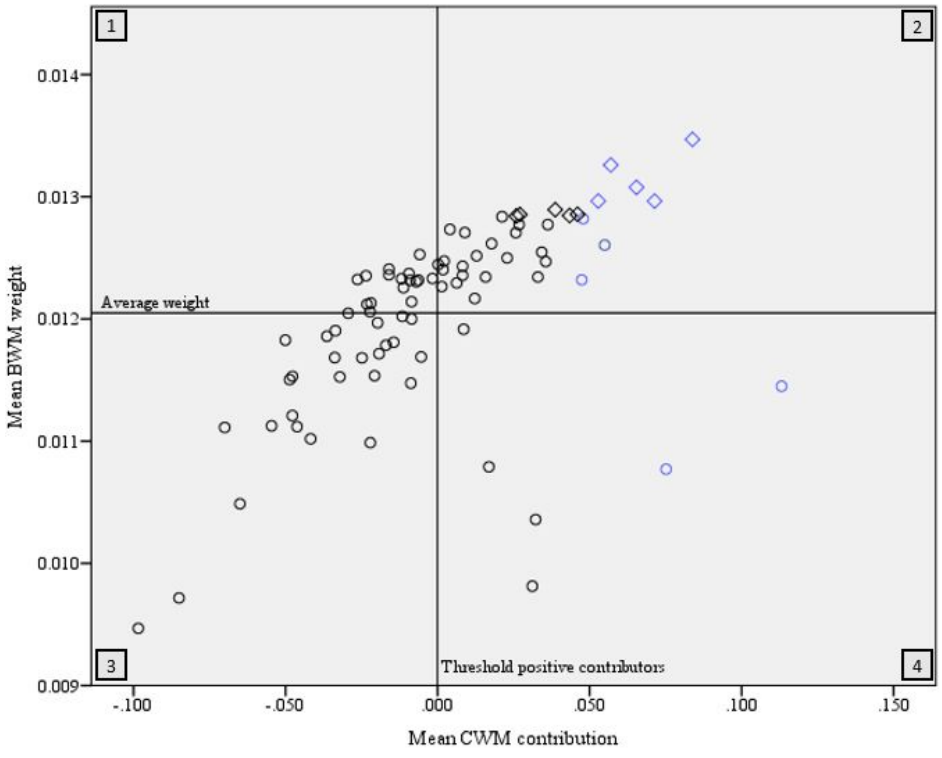
Fig. 7 Mean probability estimates for low anchor vs. high anchor
The graph shows the mean probability estimate for each question differentiated by subjects who got a low anchor vs. high anchor displayed.

Because of this difference, it is also interesting to test if the diversity of estimates for the subjects of the treatment condition is significantly lower than for those of the baseline condition. A test of equal variances shows that the null-hypothesis of equal variances can be rejected at the 5% significance level (p=0.019).[29] Hence, the variance of estimates is significantly lower in the treatment than in the baseline group. It can be negative to have less diversity, if different biases cancelled each other out before. More subjects are then required, to achieve the same level of precision (Broomell & Budescu 2009; Clemen & Winkler 1986). However, in total the positive effect of the treatment shows that the higher expertise and lower biases outweigh the disadvantages of the anchoring bias and lower diversity.

[27] The t-test test is paired by event, and the hypothesis is: H0: μ_high anchor=μ_low anchor H1: μ_high anchor>μ_low anchor. I used a parametric test since the normal distribution assumption could not be rejected at a 5% significance level using the Shapiro-Wilk test.
[28] An additional analysis showed that the anchoring effect is not significantly different between the first and second half, at a 5% significance level (p=0.15). For this, a Mann-Whitney U test was used with the hypothesis: H0: μ_first half=μ_second half H1: μ_first half>μ_second half. I used a parametric test since the normal distribution assumption could not be rejected at a 5% significance level using the Shapiro-Wilk test.
[29] For this I run an F-test of equal variance. The hypothesis is: H0: σ^2_baseline=σ^2_treatment H1:σ^2_baseline>σ^2_treatment.

4.5 Comparison of the BWM and CWM

Besides understanding the effect of the treatment, it is also helpful to understand the mechanism of the two weighted models, the BWM and the CWM better. In particular for the CWM, there is only little literature on what the model really captures. Budescu & Chen (2014) state that the CWM identifies the "experts" in the crowd. This would mean, that the positive contributors determined by the CWM are also those with individual expertise. That would show in a very similar weighting to the BWM, which determines the weighting of the subjects based on individual expertise. However, Budescu & Chen (2014) also state that the CWM selects the positive contributors based on how well they compensate for the biases of the crowd. Hence, I want to analyze if those subjects that are determined as positive contributors from the CWM are also those with individual expertise. For this, I first start by analyzing the relationship of the mean contribution of each individual $\overline{C_{ji}}$ and the mean BWM weight $\overline{q_{BWM,ji}}$, both up to the last event $i = 33$. I use the mean contribution $\overline{C_{ji}}$ and not the mean CWM weight $\overline{q_{CWM,ji}}$, since in the CWM weighting the information about the negative contributors gets lost, because the CWM weight of all negative contributors is equal to zero.



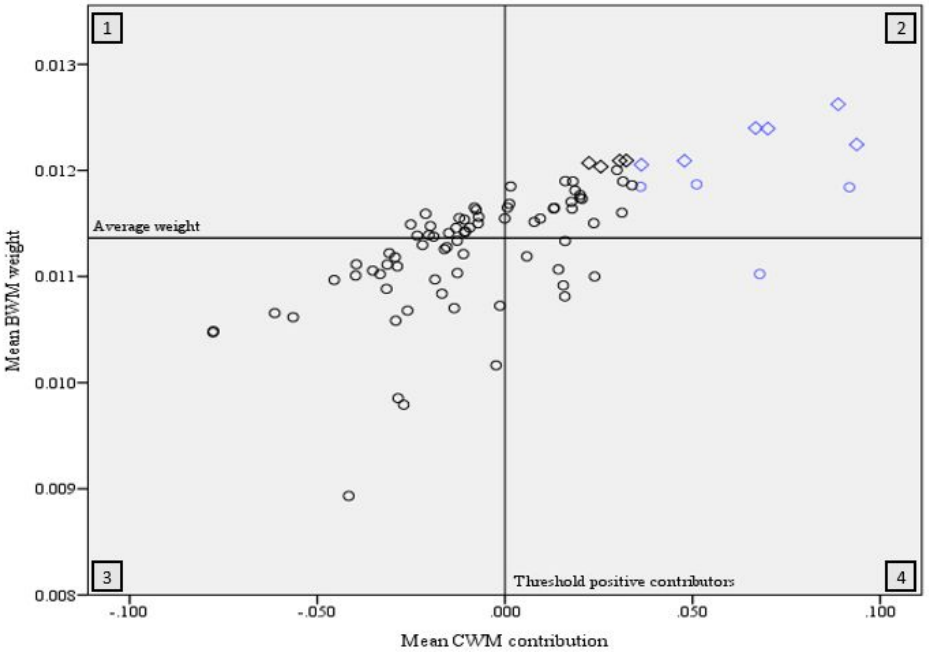


Fig. 9 Relationship of mean contribution $\overline{C_{ji}}$ and mean BWM weight $\overline{q_{BWM\,ji}}$ in the baseline (first graph) and treatment group (second graph)

The two graphs show the relationship of the mean contribution $\overline{C_{ji}}$ and mean BWM weight $\overline{q_{BWM\,ji}}$ in the baseline (first graph) and treatment group (second graph). The blue dots symbolize that this subject belongs to the Top 10 Contributors, while the squares symbolize that the subject is part of the Top 10 BWM. The horizontal line shows the threshold between above and below average BWM weights, and the vertical line shows the threshold between negative and positive contributors.

First, one can recognize a general trend in both graphs that with increasing mean contribution $\overline{C_{ji}}$ also the mean BWM weight $\overline{q_{BWM\,ji}}$ increases. This speaks for the argument that the CWM gives those people more weight who have individual expertise. In addition, it can be seen that there is in both the baseline as well as the treatment condition a large group with a mean contribution $\overline{C_{ji}}$ of around 0 and an BWM weight around the average weight. This means, that the majority of the group performs very similarly. Furthermore, in both graphs there are subjects who receive an above average weighting in the BWM but are not included in the CWM as positive contributors (area 1). This might be because the CWM has a lower number of positive contributors than there are subjects performing above average.[30] However, there are also subjects who are positive contributors in the CWM, but have a below average performance in the BWM (area 4).[31] This comes more as a surprise, since that means that the CWM also selects subjects as positive contributors who perform worse than average and also

[30] There are 40 positive contributors for the CWM in both the baseline and treatment condition. In contrast, there are 50 subjects who receive a BWM weight above average (symbolizing that they perform better than average) for the baseline condition and 51 subjects in the treatment condition.

[31] In the baseline condition, 15% of the positive contributors perform worse than average, and for the treatment condition 12.5% perform worse than average.

worse than subjects who did not get selected. Hence, not all positive contributors in the CWM can be described as having above average individual expertise. That means, that different than stated in Budescu & Chen (2014), the CWM the subjects determined as positive contributors are not necessarily "experts". Hence, there is support for Hypothesis 4 which stated that those subjects that receive a high weighting in the CWM are not the same that receive a high weighting in the BWM. To further understand this, it is also interesting to compare the mean individual scores of the Top 10 positive contributors to the Top 10 BWM, so the 10 subjects that performed best on average.

Tab. 8 Comparison of mean individual scores between Top 10 BWM and Top 10 CWM
The table compares the mean individual scores (calculated based on the 32 events) between the subjects determined as Top 10 BWM and Top 10 CWM. For this it compares the minimum, maximum, mean and median well as the standard deviation of those scores. It also differentiates between the baseline and treatment condition.

Group	Condition	n	Mean	Q25%	Median	Q75%	SD
Top 10 BWM[32]	Baseline	10	76.87	76.11	76.43	77.55	1.27
Top 10 BWM	Treatment	10	**78.61**	**78.01**	**78.29**	**79.54**	**1.31**
Top 10 CWM[33]	Baseline	10	74.50	72.62	76.16	77.55	4.59
Top 10 CWM	Treatment	10	77.15	76.00	77.49	79.54	3.15

* = p<0.1, ** = p<0.05 *** = p<0.01.

Table 8 shows that the individual mean scores are lower for the Top 10 CWM than for the Top 10 BWM, although not at a significant level.[34] The reason for this gets clear in Fig. 9. In total, only 5 of the Top 10 CWM are also in the Top 10 BWM in the baseline condition, and only 6 of the Top 10 CWM are also in the Top 10 BWM in the treatment condition. In addition, for the treatment group, the subject with the fifth-highest contribution (and hence part of the Top 10 CWM) performs even below average. For the baseline group an even stronger effect is observable, both the subject with the highest and the third-highest contribution perform below average. When looking closer at the three subjects that are determined as positive contributors by the CWM but perform below average, one can see that

[32] This group consists of the 10 subjects of a condition (baseline or treatment) that received the highest weighting at the BWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) BWM".

[33] This group consists of the 10 subjects of a condition (baseline or treatment) that received the highest weighting at the CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) CWM".

[34] The null hypothesis of a normal distribution was rejected at 5% significance level using a Shapiro-Wilk test, so I decided to use a non-parametric test. For that I used an Independent Samples Mann-Whitney U test, with the hypothesis: H0: μ_BWM=μ_CWM H1: μ_BWM<μ_CWM. The null-hypothesis could not be rejected at a 5% significance level (p_baseline =0.14, p_treatment=0.41). I used the Mann-Whitney U test because there was no pairing possible.

they tend to make extreme probability estimates. Especially the subject determined by the CWM as highest positive contributor for the baseline group only used probability estimates of either 0% or 100%. Of those probability estimates, 68.75% (22 out of 32 events) turned out to be correctly forecasted. The BWM punishes the subject for those 10 events that he did not forecast correctly, so that overall he performs below average. This is done through the squared distance to the true result. In the CWM on the other hand, subjects that "gamble" like this can perform well. It rewards subjects highly when they are correct and the crowd is not. On the other hand if the crowd is wrong, and a subject has even more extreme estimates, the negative contribution is comparably low. Unless the subject estimates for an extreme which turn out to be wrong and the majority of the crowd is correct, he would be considered as a positive contributor. This effect is also reinforced in Table 5, where one can see that the Top Contributors clearly estimate more extreme than the Top 10 performers of the BWM. Hence, an important difference between the BWM and CWM appears to be that in the CWM, subjects get more rewarded if they perform surprisingly well and get punished less if they perform (slightly) worse than the crowd. This also shows, that the CWM is more vulnerable to subjects who are just lucky. This is why for the CWM, having a large track record (with the same subjects) is even more important than for the BWM.

5 CONCLUDING REMARKS

5.1 Limitations and recommendations for further research

In this study, there have been a number of limitations that I would like to discuss, especially with concern to the experimental setting.

First, there are limitations concerning the treatment of displaying three estimates of previous subjects. I only used the estimates of 10 subjects, since it was not possible for me technically to automate this process and dynamically involve all subjects. Although there are 720 possible combinations out of those 10 subjects, it is still only a combination from 10 opinions. Hence, if possible I recommend for further research to include the estimates of all previous subjects. Furthermore, the estimates were always provided by the same three subjects over all questions. For further research however it would also be interesting to test for a difference if the subjects instead change over time.

Second, the scope of the experiment has limitations. As stated in 3.1, I expected subjects to take around 15 minutes for the questionnaire. However, the analysis of the duration shows that the median is only 9 minutes.[35] Thus, it might be that the subjects did not put enough effort into making good forecasts. The CWM is the model that would be the least affected by that, as long as there are some subjects (with high expertise) who put in enough effort, while the ones performing badly would not be taken into account. However, the other models would be harmed by this. Instead, the CWM and BWM take damage by the next limitation, while it means no harm for the other two (unweighted) models. For the experiment, there were only 32 events used, while the BWM and even more the CWM get their power over time (e.g. in Chen et al. (2016) more than 200 events were used). In addition, the scope of the experiment causes another limitation. It could be that there were too many questions for the subjects, so that they might still focus in the beginning but then lose concentration. Since the questions were not randomized to simulate a timely sorted questioning, it might be that subjects perform worse at the later questions. As especially the CWM gains its power over time, this might be a limitation that restricts the accuracy of this model more than the other models.[36]

[35] Since there were some outliers (some subjects took several hours to complete the experiment), I decided to use the median instead of the average.

[36] However, in Table B.2 in Appendix B.2, for neither the baseline nor the treatment scores there is a trend across the questions recognizable.

Third, in section 4.4.2 one can see that the mode of all estimates was at 50%, and in addition half of the estimates are below and the other half above 50%. Thus, it seems as if subjects assumed that the natural chance is 50:50. However, the natural chance of the event happening (winning) is only 33%, while it is 67% to not happen (losing or tie). Although it was stated in every question that not winning consists of both losing and tie (see Fig.1 & Fig. 2), it might be that subjects did not pay enough attention to it. In combination with Fig. 5, where the calibration curves of both conditions are mainly below the identity curves, this shows that subjects in the experiment on average gave too high estimates. The effect of this limitation does not harm any model in particular, however it decreases the scores of all models. Hence, I recommend for further research to use matches were there is no tie possible as an outcome, such as world cup matches.

Fourth, there are also limitations in the area of incentives, such as that they might be considered as too low. As there was a chance to win 30€, across the 181 subjects the expected value is only 0.17€ (assuming that everyone performs equally well). In contrast, for the experiment from Chen et al. (2016), the subjects received 250$ at the end of one year if they answered at least 25 questions, regardless of their accuracy. In my study, I expect that subjects instead participated because of intrinsic motivation (helping a student) and also because of fun (I got the feedback that the task was considered fun). Intrinsic motivation can on the one hand provide enough incentives for subjects to perform well. On the other hand however, the control of the researcher is also lower (see five precepts of Smith 1976).

Fifth, the scoring rule used to determine the scores of the different models can have an impact on the results. In my thesis, same as in Budescu & Chen (2014) and Chen et al. (2016), I am using an the Brier score. However there are also other possibilities, such as a logarithmic scoring rule. In Budescu & Chen (2014), they found no change of the results when using the logarithmic scoring rule, so I also do not expect any changes for my results.

Last, due to skewness of the score distribution as well as a small sample size, I use nonparametric tests for many parts of my analysis. Although those tests have the advantage of not requiring a normal distribution, they are also more conservative than parametric tests. This should be taken into account for the results presented.

5.2 Conclusion and discussion

When a forecast for a specific event is required, one option is to make use of the wisdom of the crowd and ask a group of individuals for their opinions. This paper focused on the mathematical aggregation of different judgements using four different models: (1) the mean and (2) median as well as (3) the Brier Weighted Model and (4) the Contribution Weighted Model. The goal of this thesis was to improve those four models using two different approaches.

First, the BWM and CWM were modified by decreasing the power of past performance on the weighting for the first 10 events. It was found that while the overall prediction accuracy remained unchanged, both models became more stable with a lower standard deviation, although not significantly. This resulted in lower scores on the upper end, but also higher scores on the lower end. In addition, forecasts can be made from the very beginning. Hence, those models can be seen as a good alternative to pure unweighted models when the track record available is not large enough.

Second, a treatment was introduced to decrease biases and increase expertise. For that, half of the subjects received additional information about the estimates of three randomly selected prior subjects. It was shown that the prediction accuracy was significantly improved by the treatment for three of the four models (mean and median as well as BWM), while it was not significant for the CWM. Betting odds are a good and easy alternative for the wisdom of the crowd. However, they are not available for many events. In my study for example from the 32 events only for 14 events betting odds were available. The prediction accuracy of the four models (in the treatment condition) was not significantly different to betting odds. Hence, the proposed models in my thesis represent good alternatives for those cases.

Third, the effect of the treatment was analyzed in more detail. It was found that the treatment led to overall less low scores. On an individual level, the mean scores seemed to have improved for all groups of expertise, but especially for those with little expertise. It was further shown that the subjects from both the baseline and treatment condition were biased to make too extreme probability estimates. The treatment had the effect of decreasing this bias, by having fewer estimates of the extreme values 0% and 100%. In addition, there were distinctly more estimates of 50%. Hence, it seems as if the treatment had the effect of making

the subjects doubt themselves more, which is reinforced by a longer duration for completing the questionnaire for the subjects in the treatment condition than in the baseline condition. This doubting effect also shows up in the results of the analysis of a possible optimism bias. For both Dutch and German fans it was found that the difference between the estimates of fans and non-fans decreased when three estimates of other subjects were displayed, although not significantly. Besides increasing the expertise and making the subjects have more doubts about their answer, the treatment also has a significant anchoring effect, although debiasing techniques were applied. In total however, the effect of the treatment on the prediction accuracy is still positive. This suggests that the increase in expertise and decrease in the optimism bias and bias to make too extreme probability estimates outweigh the increase in the anchoring effect.

Last, the BWM and CWM were compared in more detail. It was shown that those subjects that are identified as improving the crowd's expertise not necessarily also possess high individual expertise. Second, it appeared as if reporting too extreme probability estimates is less punished by the CWM than by the BWM.

When I compare my findings to those of Budescu & Chen (2014) and Chen et al. (2016), there are some differences that I would like to highlight. First, both in Budescu & Chen (2014) and Chen et al. (2016), the authors find that the Contribution Weighted Model performs better than the BWM and mean model (the median model was not part of their analysis), independent of the condition. However, one major difference between my thesis and those studies besides the lower number of events is the number of subjects. Although the number is high enough according to the rule of thumb of Wagner & Suh (2014), it might be too low for the Contribution Weighted Model. In one of their studies, Budescu & Chen (2014) for example have 420 subjects, and identify 220 as positive contributors. Although the ratio of positive contributors fits with my study, the average number of positive contributors was only 44. Hence, the group of positive contributors in the CWM in my thesis might be too small to perform better than the unweighted models. Further research should look more into the necessary minimum required number of subjects for the Contribution Weighted Model.

For further research, it would also be interesting to test different specifications of the treatment, such as a variation of the number of presented estimates. In addition, Chen et al. (2016) also tested for more specifications of the CWM, such as only including positive contributors who contributed more than a certain threshold. This led to a better performance,

but was not tested in this thesis. It would be interesting to see in further research whether the effect of the two measures (modification and displaying three estimates) would be the same for a different threshold.

As for practical implications, the results must be also be considered in the context of the limitations discussed in 5.1. In total, the thesis showed how to improve forecasting models in an easy and low-cost manner, and contributed to the better understanding of the Contribution Weighted Model.

REFERENCES

Andersson, P., Edman, J., & Ekman, M. (2005). Predicting the World Cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting,* 21, 565–576.

Angrist, J. & Pischke, J.S. (2009). *Mostly Harmless Econometrics*. Princeton and Oxford: Princeton University Press.

Ariely, D., Loewenstein, G., Prelec; D. (2003). Coherent Arbitrariness: Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics*, 118(1):73–106.

Armstrong, J.S. (2001). Combining forecasts. In Armstrong, J.S., *Principles of forecasting: a handbook for researchers and practitioners*, Kluwer Academic Publishing, 2001, p. 417-439.

Aronson, E., Wilson, T.D., & Akert, R. (2010). *Social psychology*. Upper Saddle River: Prentice Hall.

Arrow, K.J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J.O., Levmore, S., Litan, R., Milgrom, P., Nelson, F.D., Neumann, G.R., Ottaviani, M., Schelling, T.C., Shiller, R.J., Smith, V.L., Snowberg, E., Sunstein, C.R., Tetlock, P.C., Tetlock, P.E., Varian, H.R., Wolfers, J., Zitzewitz, E. (2008). Economics. The promise of prediction markets. *Science*, 320(5878):877-8.

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

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C. & Sugden, R. (2009). *Experimental Economics: Rethinking the Rules*, Princeton University Press.

Baltussen, G., Post, T., van den Assem, M. & Wakker, P. (2012). Random Incentive Systems in a Dynamic Choice Experiment, *Experimental Economics*, 15(3):418-443.

BBC (2018). International Friendlies Scores & Fixtures. *BBC*, Retrieved at April, 2 from the World Wide Web: https://www.bbc.co.uk/sport/football/international-friendlies/scores-fixtures/2018-05?filter=fixtures.

Block, R. & Harper, D. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49(2):188-207.

Budescu, D. & Chen, E. (2014). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science,* 61(2):267-280.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1):1-3.

Broomell, S.B. & Budescu, D.V. (2009). Why Are Experts Correlated? Decomposing Correlations Between Judges. *Psychometrika,* 74:531-551.

Cape, P. (2009). *Slider scales in online surveys.* Paper presented at the CASRO Panel Conference, February 2-3, 2009, New Orleans, LA. Retrieved May 18, 2018 from the World Wide Web:http://www.surveysampling.com/ssi-media/Corporate/white_papers/SSI-Sliders-White-Pape.image.

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

Clemen, R.T., & Winkler, R.L. (1986). Combining economic forecasts. *Journal of Business and Economic Statistics*, 4:39–46.

Cooke, R.M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford Univ. Press.

Davis-Stober, C., Budescu, D., Dana, J., Broomell, S.B. (2014). When is a crowd wise?. *Behavior Research Methods,* 46(1).

Epley, N. & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, 12:391-396.

Forrest, D., Goddard, J., Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting,* 21:551– 564.

French., S. (2011). Expert judgement, meta-analysis and participatory risk analysis. *Decision Analysis,* 9(2):119-127.
Galton, F. (1907-1). "Vox Populi". *Nature*, 75 (1952), 450-1.
Galton, F. (1907-2). "Letters to the Editor: The Ballot-Box". *Nature*. 75 (1952), 450-1.
Gilovich T., Griffin D.W., Kahneman D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge: Cambridge University Press.
Gröschner, C., & Raab, M. (2006). Vorhersagen im Fußball: Deskriptive und normative Aspekte von Vorhersagemodellen im Sport [Forecasting soccer: Descriptive and normative aspects of forecasting models in sports]. *Zeitschrift für Sportpsychologie, 13*, 23–36.
Hahn, R.W. & Tetlock, P.C. (2006). A New Approach for Regulating Information Markets. *Journal of Regulatory Economics*, 29:265-281.
Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, 112:494–508.
Herzog, S.M. & Hertwig, R. (2011). The wisdom of ignorant crowds: predicting sport outcomes by mere recognition. *Judgment and decision making*, 6:58-72.
Hvattum, L.M. & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football, *International Journal of Forecasting*, 26:460-470.
Jain, K., Bearden, J.N., Filipowicz, A. (2011). Diverse personalities make the crowd wiser: How personality can affect the accuracy of aggregated judgements. (working paper).
Kahneman D, Slovic P, Tversky A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Knight, Frank H. (1921). Risk, Uncertainty, and Profit. *Library of Economics and Liberty*. Retrieved May 14, 2018 from the World Wide Web: http://www.econlib.org/library/Knight/knRUP.html
Koehler, D.J., & Beauregard, T.A. (2006). Illusion of confirmation from exposure to another's hypothesis. *Journal of Behavioral Decision Making*, 19:61-78.
Larrick, R., Soll, J. (2006). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle. *Management Science,* 52(1):111-127.
Linstone, H.A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Reading, Mass: Addison-Wesley Pub. Co., Advanced Book Program.
Lorge, I., Fox, D., Davitz, J. & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological Bulletin* 55(6):337-72.
Lyon, A., Pacuit E. (2013). *The Wisdom of Crowds: Methods of Human Judgement Aggregation*. In: Michelucci P. Handbook of Human Computation. Springer, New York, NY.
Mannes, A.E., Soll, J.B., Larrick, R.P. (2014). The wisdom of selected crowds. *J. Personality Soc. Psych.,*107(2):276–299.
Manski, C. (2006). Interpreting the predictions of prediction markets. *Economics Letters*, 91(3):425-429.
Moore, D.A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2):502-517.
Nielsen Sports (2018). Global interest in football, *Nielson Sports*, Retrieved May 18, 2018 from the World Wide Web: http://nielsensports.com/global-interest-football/.
Pachur, T., & Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica,* 125: 99–116.
Peeters, T.L.P.R. (2018). Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, 34(1):17–29.
Polymath, D.H.J. (2012). A new proof of the density Hales-Jewett theorem. *Annals of Mathematics*, 175(3): 1283-1327.
Rowe and Wright (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, 15(4).

Shanteau, J., & Stewart, T.R. (1992). Why study expert decision making? Some historical perspectives and comments. *Organizational Behavior and Human Decision Processes*,53(2), 95.

Simmons, J., Nelson, L.D., Galak, J. & Frederick, S. (2011). Intuitive biases in choice vs. estimation: Implications for the wisdom of crowd*., J. Consumer Res*., 38(1):1–15.

Sniezek, J.A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43*(1):1-28.

Soll, J.B., & Larrick, R.P. (2009). Strategies for revising judgment: how (and how well) people use others' opinions, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35:780–805.

Song, C., Boulier, B.L., Stekler, H.O. (2007). The comparative accuracy of judgmental and model forecasts of american football games. *International Journal of Forecasting,* 23:405–413.

Štrumbelj, E. (2014). On Determining Probability Forecasts from Betting Odds. *International Journal of Forecasting* 30(4):934–943.

Sunstein, C. (2006). *Infotopia: How Many Minds Produce Knowledge*. Oxford, New York: Oxford University Press.

Surowiecki J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* New York: Anchor books.

Taylor, D.M. & Doria, J.R. (1981). Self-serving and group-serving bias in attribution. *Journal of Social Psychology*. 113 (2):201–211.

Tetlock, P.E. (2005). *Expert political judgment: How good is it? how can we know?*. Princeton, N.J.: Princeton University Press.

Thaler, R.H. & Ziemba, W.T. (1988). Anomalies: Parimutuel Betting Markets: Racetracks and Lotteries. *Journal of Economic Perspectives*, 2 (2): 161-174.

Treynor J.L. (1987). Market Efficiency and the Bean Jar Experiment, *Financial Analysts Journal,* 43 (3), 50-53.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124-1131.

von Winterfeldt, D. (1999). On the relevance of behavioral decision research for decision analysis. Pp. 133–154 in Shanteau, J., Mellers, B.A., Schum, D.A. (eds). *Decision Science and Technology: Reflections on the Contributions of Ward Edwards*. Norwell: Kluwer.

Wagner, C. & Suh, A. (2014). The Wisdom of Crowds: Impact of Collective Size and Expertise Transfer on Collective Performance. *47th Hawaii International Conference on System Sciences*, Waikoloa, HI, 594-603.

Whyte, G. & Sebenius, J. (1997). The Effect of Multiple Anchors on Anchoring in Individual and Group Judgment. *Organizational Behavior and Human Decision Processes*, 69(1):75-85.

Wolfers, J. & Zitzewitz, E. (2006). Interpreting Prediction Market Prices as Probabilities. *NBER Working Paper,* No. w12200.

Young, R.M.B. (2010). Decomposition of the Brier score for weighted forecast-verification pairs. *Quarterly Journal of the Royal Meteorological Society*, 136(650):1364–1370.

APPENDIX

A. Survey questions

Tab. A.1 Survey questions introduction

Question	Answer possibilities
Dear subject, Thank you very much for helping me with my Master Thesis! In the following, you'll be asked to forecast the probability of a football team winning a match. Because it's almost time for the FIFA World Cup 2018, you'll be asked to forecast some of the last friendly matches before the World Cup. There will be some demographics questions first, before starting with the forecasting questions. Once more thanks a lot! :-) Julia	None

Tab. A.2 Survey questions demographics

Question	Answer possibilities
What is your age?	[Text entry]
What is your gender?	● Female ● Male ● Other
What is your current employment status?	● Employed full time ● Employed part time ● Unemployed looking for work ● Unemployed not looking for ● work ● Retired ● Student ● Disabled ● Other
How often do you watch football matches of national football teams (countries)?	● I watch every national football match possible ● I sometimes watch national football matches (e.g. only during the FIFA World Cup) ● I never watch national football matches
What is the national football team you like most?	[List of all countries]
Thank you for answering the demographic questions, you are now starting with the forecasting task!	None

Tab. A.3 Survey questions baseline group

Question	Answer possibilities
Baseline group: Rules of the tournament:	[Text entry]

In the following, you are asked to give an estimate for the chance that a football team wins a match.

For every estimate, you can earn up to 10.000 points. The points are given based on how close you are to the true result. If the team wins, the real result is 100, and if the result is draw or losing, the true result is 0.
(More information to the rule I'm using: (100 points - distance from true result)².
Example: If you estimate that the chance of Germany winning against Japan is 80% and Germany wins, you get 6400 points (100 - 20 distance)²=80². If Germany does not win (either lose or draw), you get 400 points (100 points-80 points distance)².)

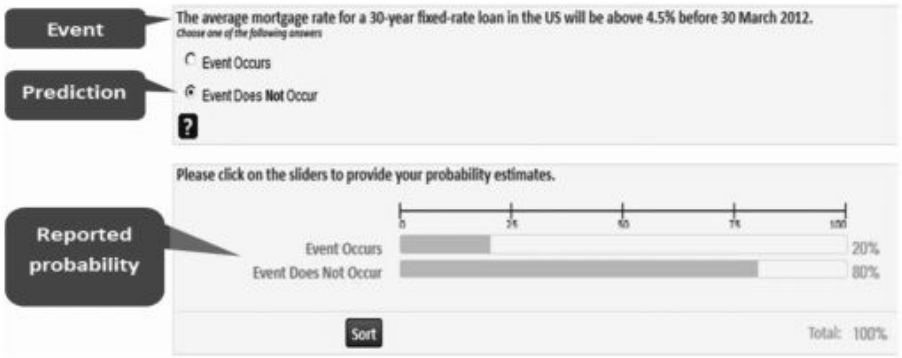
Of all subjects, one subject will win 30€. The winner is determined by a lottery. The more points you collect in the forecasting task, the more lottery tickets you get, so the higher your chance to win.
In order to inform the winner, please provide me with your e-mail address below. Of course, if you don't want to participate in the tournament for the prize, you don't need to provide me with your e-mail address.

Please estimate the chances of the following team winning on May 26th (=not losing and no draw): China PR winning against Jordan	[Slider from 0% to 100%]
[same question for all matches with changing dates]: Kuwait winning against Egypt	[Slider from 0% to 100%]
Iran winning against Turkey South Korea winning against Honduras	[Slider from 0% to 100%]
Bosnia-Herzegovina winning against Montenegro	[Slider from 0% to 100%]
France winning against Republic of Ireland	[Slider from 0% to 100%]
Italy winning against Saudi Arabia	[Slider from 0% to 100%]
Nigeria winning against DR Kongo	[Slider from 0% to 100%]
Portugal winning against Tunisia	[Slider from 0% to 100%]
USA winning against Bolivia	[Slider from 0% to 100%]
Mexico winning against Wales	[Slider from 0% to 100%]
Malta winning against Armenia	[Slider from 0% to 100%]
Azerbaijan winning against Kyrgyzstan	[Slider from 0% to 100%]
Panama winning against Northern Ireland	[Slider from 0% to 100%]
Peru winning against Scotland	[Slider from 0% to 100%]
Japan winning against Ghana	[Slider from 0% to 100%]
Morocco winning against Ukraine	[Slider from 0% to 100%]
Austria winning against Russia	[Slider from 0% to 100%]
Luxembourg winning against Senegal	[Slider from 0% to 100%]
Slovakia winning against Netherlands	[Slider from 0% to 100%]
South Korea winning against Bosnia-Herzegovina	[Slider from 0% to 100%]

Georgia winning against Malta	[Slider from 0% to 100%]
Tunisia winning against Turkey	[Slider from 0% to 100%]
Australia winning against Czech Republic	[Slider from 0% to 100%]
France winning against Italy	[Slider from 0% to 100%]
England winning against Nigeria	[Slider from 0% to 100%]
Montenegro winning against Slovenia	[Slider from 0% to 100%]
Sweden winning against Denmark	[Slider from 0% to 100%]
Republic of Ireland winning against USA	[Slider from 0% to 100%]
Austria winning against Germany	[Slider from 0% to 100%]
Belgium winning against Portugal	[Slider from 0% to 100%]
Iceland winning against Norway	[Slider from 0% to 100%]
Mexico winning against Scotland	[Slider from 0% to 100%]

Tab. A.4 Survey questions treatment group

Question	Answer possibilities
Rules of the tournament: In the following, you are asked to give an estimate for the chance that a football team wins a match. As an orientation, you will always see on top the estimates of three other random subjects. Those three subjects are going to be the same across all questions. For every estimate, you can earn up to 10.000 points. The points are given based on how close you are to the true result. If the team wins, the real result is 100, and if the result is draw or losing, the true result is 0. (More information to the rule I'm using: (100 points - distance from true result)². Example: 1. If you estimate that the chance of Germany winning against Japan is 80% and Germany wins, you get 6400 points (100 - 20 distance)²=80². If Germany does not win (either lose or draw), you get 400 points (100 points-80 points distance)².) Of all subjects, one subject will win 30€. The winner is determined by a lottery. The more points you collect in the forecasting task, the more lottery tickets you get, so the higher your chance to win. In order to inform the winner, please provide me with your e-mail address below. Of course, if you don't want to participate in the tournament for the prize, you don't need to provide me with your e-mail address.	[Text entry]
Please estimate the chances of the following team winning on May 26th (=not losing and no draw). 3 other subjects estimated the chance of China PR winning against Jordan A: [x]%, B: [y]%, C: [z]%	[Slider from 0% to 100%]
[same events as for baseline group]	[Slider from 0% to 100%]



Source. http://forecastingace.com (site discontinued).

	Probabilities of euro area inflation* Year-on-year change in the HICP					
	2013	**2014**	**2015**	**December 2013**	**December 2014**	**5 years ahead (2017)**
< -1.0%						
-1.0- -0.6%						
-0.5- -0.1%						
0.0-0.4%						
0.5-0.9%						
1.0-1.4%						
1.5-1.9%						
2.0-2.4%						
2.5-2.9%						
3.0-3.4%						
3.5-3.9%						
≥ 4.0%						
Total	100	100	100	100	100	100

* Defined on the basis of the Harmonised Index of Consumer Prices produced by Eurostat. Probabilities should sum to 100%. Average of the period.

Fig. A.1 Examples for questions in Budescu & Chen (2014)
The first figure shows the question format of Budescu & Chen's (2014) first approach and the second figure shows the question format their second approach (using ECB data), as explained in 3.1.

B. Additional analyses

B.1 Robustness: Modified version across 15 and 20 events

As a robustness check, I decrease the power of the weights of the BWM/CWM not only for the first 10 but instead 15 and 20 events. The results are very similar: Especially the 3 low scores are improved (or the same), while the 3 highest scores are lower (or the same) for the modified version. The mean of the modified models is higher than those of the unmodified ones, however they are not significantly different at a 5% significance level ($p_{BWM,15} = 0.58$, $p_{BWM,20} = 0.47$, $p_{CWM,15} = 0.20$, $p_{CWM,20} = 0.20$).[37] The same holds for the variance: While the variance is lower for the modified versions, the difference is not significant at a 5% significance level ($p_{BWM,15} = 0.50$, $p_{BWM,20} = 0.50$, $p_{CWM,15} = 0.42$, $p_{CWM,20} = 0.38$).[38]

[37] For this, I run a paired-sample Wilcoxon test (paired by event), with the following hypothesis:
H0: μ_ unmodified=μ_modified H1: μ_ unmodified<μ_modified.
I used a non-parametric test because there are not enough observations to assume a normal distribution of the scores. In addition, the null-hypothesis of normal distribution was rejected at a 5% significance level for all models except one using a Shapiro-Wilk test. Because of consistency reasons, I used a non-parametric test.

[38] The hypothesis for the F-test is: H0: σ^2_modified=σ^2_ unmodified H1:σ^2_modified<σ^2_ unmodified.

Tab. B.1 Effect of modification on accuracy of models - robustness check with 15 and 20 events
This table shows the results of a robustness test to the analysis done in chapter 4.4.1, by decreasing the power of the weights of the BWM/CWM not only for the first 10 but instead 15 and 20 events. The table compares the scores of the BWM and CWM in the modified and unmodified version. This is across the first 14/19 events, where the weighting in the modified version is not fully determined by the BWM or CWM yet. For the comparison of the standard deviation (SD), the first event is excluded, to make the modified and unmodified versions comparable.

Event	BWM unmodified		BWM modified		CWM unmodified		CWM modified	
	I=14	I=19	ω = 1 at i = 15	ω = 1 at i = 20	I=14	I=19	ω = 1 at i = 15	ω = 1 at i = 20
E1			89.5↑	89.5↑			89.5↑	89.5↑
E2	91.8	91.8	91.3↓	91.3↓	94.7	94.7	91.5↓	91.5↓
E3	87.2	87.2	87.0↓	87.0↓	85.8	85.8	86.8↑	86.8↑
E4	63.9	63.9	64.1↑	64.1↑	58.2	58.2	62.9↑	63.2↑
E5	97.0	97.0	96.9↓	96.9↓	97.8	97.8	97.2↓	97.1↓
E6	98.6	98.6	98.6→	98.6→	99.5	99.5	98.9↓	98.8↓
E7	53.4	53.4	53.6↑	53.7↑	49.4	49.4	51.9↑	52.4↑
E8	21.9	21.9	21.9→	21.9→	17.6	17.6	19.8↑	20.4↑
E9	89.9	89.9	90.1↑	90.1↑	88.3	88.3	89.1↑	89.3↑
E10	67.2	67.2	66.9↓	66.9↓	72.9	72.9	69.2↓	68.6↓
E11	83.5	83.5	83.8↑	83.8↑	80.9	80.9	82.0↑	82.5↑
E12	66.1	66.1	66.0↓	66.1→	65.2	65.2	69.3↑	68.5↑
E13	88.7	88.7	88.9↑	88.9↑	87.8	87.8	87.8→	88.1↑
E14	71.1	71.1	71.0	71.0↓	64.9	64.9	67.3↑	68.5↑
E15		72.3		72.4↑		72.9		70.9↓
E16		83.1		83.1→		85.7		82.9↓
E17		73.2		73.2→		71.8		73.0↑
E18		87.0		87.0→		87.0		81.7↓
E19		92.9		93.0↑		95.3		93.8↓
Mean	75.4	77.2	76.4↑	77.8↑	74.1	76.4	76.0↑	77.2↑
SD	21.4	18.8	21.4→	18.7↓	23.1	20.4	21.9↓	18.9↓

B.2 Additional analyses of effect of displaying three estimates

When analyzing the effect of the treatment on the unmodified BWM and CWM, there are some results worth mentioning with regard to the CWM. First, while the scores of the treatment condition for the modified version were not significantly higher than those of the baseline condition, this is the case for the unmodified version for the first 10 events. Second, there is one decrease that requires some additional explanation, which is the decrease of the minimum of the CWM from 19.6 to 15.6. The question was here to estimate the chance of Portugal winning against Tunisia. When looking at the estimates, one can see that the subjects of the treatment group estimated the chance of Portugal to win against Tunisia higher than those of the baseline group. Although one can discuss what the objective probability would have been, this match ended in a tie, which made the subjects of the treatment group perform worse than those of the baseline group. That is because the CWM requires a large track record in order to be able to successfully select its positive contributors. Since this minimum was at the eighth event, this requirement was not fulfilled yet, so that the selection of the positive contributors that determined the forecast of this event was based on only seven other events, possibly with little correlation. This gives a good reason on why introducing the modification of the BWM and CWM makes them more stable.

Tab. B.2.1 Comparison of the prediction accuracy of all models
In addition to Tab. 3 in chapter 4.4.2, this table also shows an overview of the scores of the unmodified BWM and CWM both in the baseline and treatment condition. The scores are across all 32 events. The p-value in the last column report if the scores in the treatment condition are significantly higher than in the baseline condition (for both BWM and CWM).

Model	Condition	Min	Max	Mean	Median	SD	p[39]
BWM unmodified	Baseline	21.0	98.7	76.1	81.3	17.6	
	Treatment	22.6	98.6	76.8	80.8	17.5	0.04**
CWM unmodified	Baseline	19.6	99.3	74.8	79.1	19.8	
	Treatment	15.6	**99.5**	75.9	**82.4**	19.8	0.03**

* = p<0.1, ** = p<0.05 *** = p<0.01; n=31

[39] For both the BWM and CWM, I am testing the following hypothesis: H0: μ_baseline=μ_treatment H1: μ_baseline<μ_treatment by using a paired-sample Wilcoxon test, comparing the scores paired by event. I used a nonparametric test since the normality assumption of normal distribution was rejected at a 5% significance level using a Shapiro-Wilk-test.This is likely to be due to a right-skewness of the data.

To understand the effect of the treatment as well as the models better, a boxplot is a good way of depicting the distribution of the scores of the models. For this purpose, I also included the unmodified versions of the two weighted models on the right side. First, when comparing the weighted and unweighted models one can see that the lower whisker is larger for the weighted models, same as the displayed outliers. However, for the interquartile range there is no difference observable. The next difference between the weighted and unweighted models is the 75% quartile, which seems higher for the unweighted models.

When comparing the modified and unmodified weighted models, there is no clear difference in the interquartile range, but instead the whiskers are larger. This also explains why the variance is lower for the modified models.

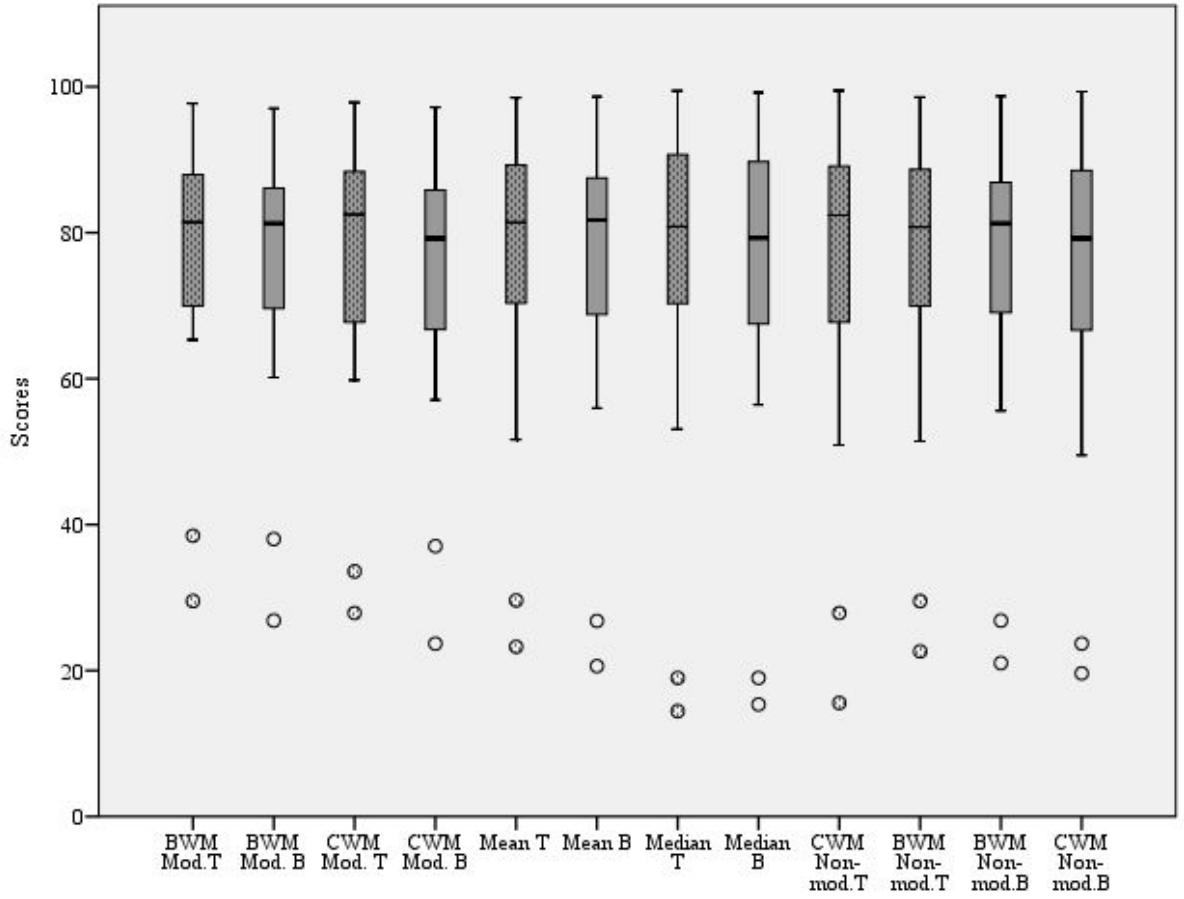


Fig. B.2.1 Boxplots of all scores of the models across the events
The figure shows the distribution of the scores of all models across the 31/32 events. For the modified BWM/CWM as well as the mean and median model, there are scores of 32 events used, while there are only 31 for the unmodified BWM/CWM. All models from the treatment condition are highlighted by points inside their IQR in the boxplots.

In order to understand the models better, it is also interesting to look in addition at the development of the scores over the events. Both for the treatment and the baseline condition

one can see that for neither of the models there is a trend visible, which one could expect from the two weighted models that benefit from having a larger track record. However, it might also be that 32 events are too little to recognize that.

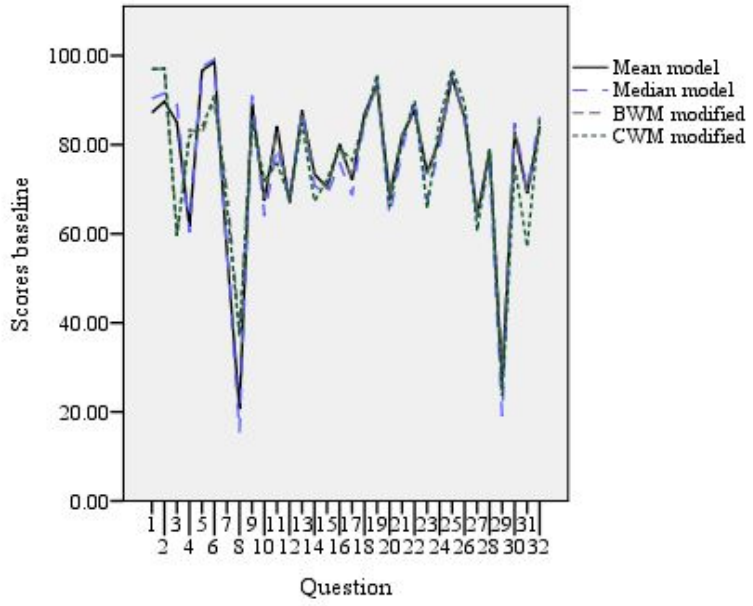


Fig. B.2.2 Development of the scores over the events for the baseline condition
The graph shows the scores of the four models for the baseline condition for each of the 32 events.

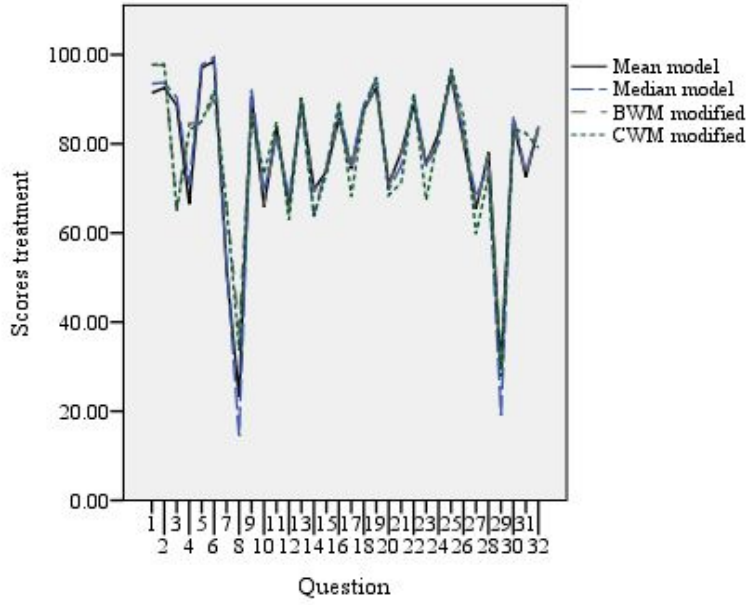


Fig. B.2.3 Development of the scores over the events for the treatment condition
The graph shows the scores of the four models for the treatment condition for each of the 32 events.

B.3 Comparison of mean individual scores on group level

As a more detailed analysis to chapter 4.4.1, the following table compares the mean individual scores on a group level. One can see that all groups except for the Top 10 positive contributors of the CWM have significantly higher (at a 5% significance level) mean scores in the treatment than in the baseline condition. Unsurprisingly, the Top BWM in the treatment condition has the highest mean individual scores, since it is the model that determines the weights based on individual expertise, as also discussed in chapter 4.5.

Tab. B.3 Comparison of mean individual scores for groups between baseline and treatment
The table compares mean individual scores (calculated based on the 32 events) of different groups of subjects. For this it compares the minimum, maximum, mean and median well as the standard deviation of those scores. It also differentiates between the baseline and treatment condition. The highest minimum, maximum, mean and median score as well as the lowest standard deviation across all groups and conditions are highlighted in bold. In addition, it shows if the difference between the baseline and treatment condition is significantly higher for each group of subjects.

Group	Condition	n	Mean	Q25%	Median	Q75%	SD	p
All subjects	Baseline	83	71.56	69.30	72.62	74.26	4.43	
All subjects	Treatment	88	73.40	71.47	74.00	75.73	3.67	0.003***
Top 10 BWM[40]	Baseline	10	76.87	76.11	76.43	77.55	1.27	
Top 10 BWM	Treatment	10	**78.61**	**78.01**	**78.29**	**79.54**	**1.31**	0.005***
Others (not Top 10) BWM	Baseline	73	70.83	68.88	72.29	73.71	4.21	
Others (not Top 10) BWM	Treatment	78	72.73	70.75	73.75	75.01	3.31	0.001***
Top 10 CWM[41]	Baseline	10	74.50	72.62	76.16	77.55	4.59	
Top 10 CWM	Treatment	10	77.15	76.00	77.49	79.54	3.15	0.075*
Others (not Top 10) CWM	Baseline	73	71.16	69.30	72.51	73.78	4.29	
Others (not Top 10) CWM	Treatment	78	72.92	71.32	73.83	75.05	3.46	0.003***
Pos. contributors[42] (CWM)	Baseline	40	73.51	73.12	74.26	76.11	4.29	
Pos.contributors (CWM)	Treatment	40	75.59	74.90	76.0	76.90	2.74	0.004***
Neg. contributors (CWM)	Baseline	43	69.75	68.19	70.89	72.51	3.77	
Neg.contributors (CWM)	Treatment	48	71.58	69.64	72.22	73.98		0.002***

* = p<0.1, ** = p<0.05 *** = p<0.01.

[40] This group consists of the 10 subjects of a condition (baseline or treatment) that received the highest weighting at the BWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) BWM".

[41] This group consists of the 10 subjects of a condition (baseline or treatment) that received the highest weighting at the CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) CWM".

[42] This group consists of those subjects of a condition (baseline or treatment) that received a positive weighting at the CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition who were excluded from the weighting of the CWM are part of the group "Negative contributors CWM".

B.4 Calibration curves by positive/negative contributors

Besides comparing the calibration curves between the baseline and treatment condition, one can also compare groups of subjects. The most interesting one is the comparison between the positive and negative contributors in the CWM. Since the positive contributors are those that are determined based on how well they compensate the biases of the rest, one might expect that the curves are mirrored at the identity line. However, one can see in the graph that this is not the case. However, there is another difference between the two groups. The positive contributors seem to be closer to the true outcomes when estimating probabilities between 0% and 10%, while the negative contributors seem to be closer when estimating probabilities between 90% and 100%.

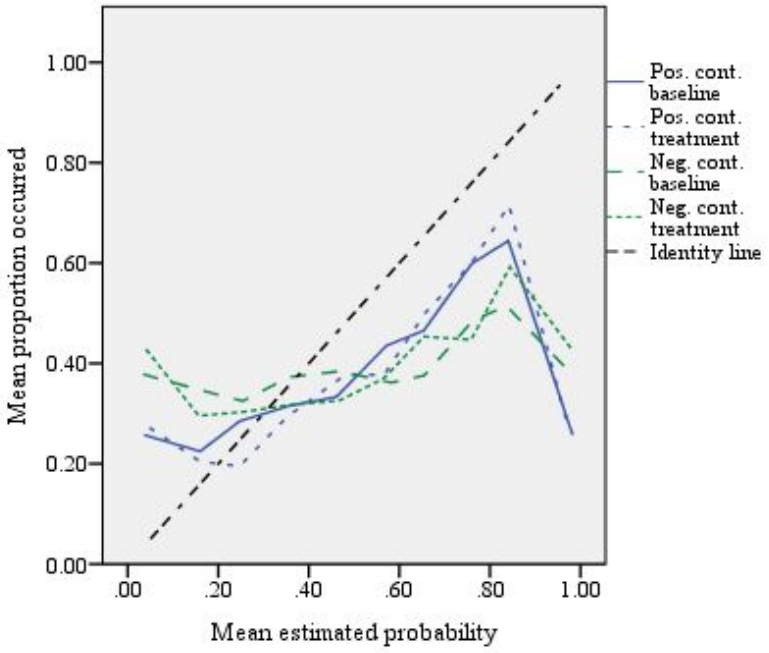


Fig. B.4 Calibration of CWM groups
This figure adds on to Fig. 5 in 4.4.2. It depicts the relationship between the mean estimated probability of a group of probability estimates (e.g. 0%-10%) for an event to occur, and the mean proportion that the event actually occurred. This relationship is represented by the calibration curves, which are shown for the positive and negative contributors of the CWM of both the baseline and treatment condition. In addition, an identity line is shown.

B.5 Estimation behavior with regard to multiples of 10%

In Figure 6 one could see that in both the baseline and treatment condition, subjects tended to provide probability estimates in multiples of 10%. Table B.5 in addition shows that this behavior is 6.2% lower for the treatment condition than the baseline condition. However, when looking at the different groups, this trend is not constant. Although the groups of Top 10 BWM, Top10 CWM and positive contributors have the lowest relative frequencies of estimating multiples of 10 in the baseline group, the Top 10 BWM and positive contributors have 44.9% and 88.4% respectively higher relative frequencies in the treatment condition. This seems counterintuitive, especially because the relative frequency of estimating multiples of 10% is lower for all other groups in the treatment condition than in the baseline condition. A possible explanation could be that those subjects who are not in one of those top-performing groups try to be too exact, while the experts do not try to be more exact than they can be. In Mellers et al. (2015), they also find that for so called superforecasters, which are experts in forecasting, rounding to the next 5 or 10 does change the performance, while it does not for all others. In addition, this could also be explained with Tab.5, where for exactly those group the frequency of estimating 50% increased immensely from the baseline to the treatment condition.

Tab. B.5 Relative frequencies of multiples of 10 as probability estimates across groups

The table shows the relative frequency that the the probability estimates predicted by a specific group of subjects were multiples of 10%. It also differentiates between baseline and treatment condition for each group of subjects.

Groups	Multiples of 10	Δ
All subjects Baseline	45.1%	
All subjects Treatment	42.3%	-6.2%
Top 10 BWM[43] Baseline	30.3%	
Top 10 BWM Treatment	43.9%	+44.9%
Others (not Top 10) BWM Baseline	47.1%	
Others (not Top 10) BWM Treatment	42.1%	-10.6%
Top 10 CWM[44] Baseline	39.7%	
Top 10 CWM Treatment	39.7%	0.0%
Others (not Top 10) CWM Baseline	45.8%	
Others (not Top 10) CWM Treatment	42.6%	-7.0%
Positive contributors[45] (CWM) Baseline	39.7%	
Positive contributors (CWM) Treatment	74.8%	+88.4%
Negative contributors (CWM) Baseline	45.8%	
Negative contributors (CWM) Treatment	30.2%	-34.1%

[43] This group consists of the 10 subjects of a condition (baseline or treatment) that received the 10 highest weightings at the (modified) BWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) BWM".

[44] This group consists of the 10 subjects of a condition (baseline or treatment) that received the 10 highest weightings at the (modified) CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition are part of the group "Others (not Top 10) CWM".

[45] This group consists of those subjects of a condition (baseline or treatment) that received a positive weighting at the (modified) CWM for the last event. This is because in this weighting, the performance of all previous events is included. All other subjects of this condition who were excluded from the weighting of the CWM are part of the group "Negative contributors CWM".