ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Behavioural Economics

Wisdom of Crowds and the Home Bias in Sports Betting Markets

Name student: Cliff van Koeverden

Student ID number: 435494

Supervisor: Benjamin Tereick

Second assessor: Tong Wang

Date final version: xxx

Abstract

In recent years, many papers have tried to improve the wisdom of crowds. However, it is still unclear which aggregation method would work best in which situation. When a large share of the crowd is biased in their predictions, most aggregation methods often give wrong predictions. The Surprisingly Popular method has shown to give promising results even when a large share of the crowd is biased. Sports betting markets have been of interest in the wisdom of crowds literature and therefore this paper looked into whether using the *Surprisingly Popular* method would improve the predictions of crowds on the outcome of sports events by controlling for the *Home bias*. This was tested through an experimental design, in which a total of 158 subjects filled in a survey containing statements on historical football matches in the Dutch football league the Eredivisie over the last ten seasons (2009-2020). The performance of the *Surprisingly Popular* method was then compared to the performance of the confidence-weighted prediction method, the most-confident prediction method and the majority prediction method. The results of this research indicate that the confidence-weighted method significantly outperforms the other aggregation methods, including the *Surprisingly Popular* method. A limitation of this paper is the use of historical data instead of predictions on future fixtures.

Introduction

In a situation in which a group of subjects must give predictions on certain outcomes, often the average prediction of this group is more accurate than estimates made by individuals in this group. Generally, this concept is called *the Wisdom of Crowds* (Mannes, Soll, & Larrick, 2014). However, this concept does not always work. In some situations, taking the average of the group gives a value close to the real answer, in other situations this strategy gives less precise estimates. For example, when a large share of the group exhibits the same bias (Kao et al., 2018), in such situations an average of the estimates will likely give a wrong prediction. Over the years, many aggregation methods have been invented and applied to estimates of crowds, in an attempt to find a method that is able to extract this wisdom from within the crowd the best. However, it is still unclear which method work best in which situations. In this paper an experimental approach has been taken to look into which aggregation method gives the best predictions.

In 1906, a weight-judging competition was held among visitors of the annual regional country fair in Plymouth. The visitors of this fair were invited to buy tickets to participate in a lottery in which they had to guess the weight of an ox. On these tickets, they had to write down their name, address, and their estimate of what they believed the ox would weight in lbs. Afterwards, these tickets were lent to Francis Galton, who claimed that the average competitor was likely to be as well fitted for making a correct estimate of the weight of the ox, as the average voter is in judging the merits of political issues (Galton, 1907). Francis Galton used the results of this weight-judging competition for his research into the *Vox Populi* (voice of the people), in which one value or estimate is used to represent a population.

A total of 787 tickets were used in this research. Francis Galton used the *Democratic* principle, in which each ticket represents one, and the same value. Out of all these tickets, he picked the median (middle value) estimate to be the *Vox Populi.* The median estimate turned out to be 1207 lbs., while the measured weight of the ox was 1198 lbs., showing how close Galton´s approximated *Vox Populi* was to the actual weight of the ox. This is a famous example of *the Wisdom of Crowds*, which is an aggregation of information within a group of people resulting in a decision, that is often better than could have been made by any of the individual members of this group (Surowiecki, 2004).

Francis Galton believed the median estimate of a crowd to best reflect the voice of this crowd. However, averaging the estimates of all 787 tickets in the weight-judging competition

mentioned before, would amount to a weight of 1197 lbs., which is an almost perfect estimate of the weight of the ox (Surowiecki, 2004), even better than the median estimate. Both these approaches show how predictions made by a group can be used to estimate a single outcome that is close to the actual value. However, it is unclear which of these aggregation methods is superior. Furthermore, it is unlikely that the almost perfectly matching weights are due to pure chance.

Mannes, Soll, and Larrick (2014), designed a new strategy to aggregate the predictions of a crowd in an attempt to improve their estimates. They created the *select-crowd strategy*, in which subjects are ranked on their ability to give good predictions. The ability of these subjects is measured based on their prediction performance in the past. By using the predictions of the five best judges within the group, they were able to improve overall estimates. They also showed that the strategy of averaging or taking the median estimate of a group is not robust.

Over the years, the concept of *the Wisdom of Crowds* has been applied in several important fields. A striking example of the application of this theory is found in election forecasting polls. Franch (2017) used this concept to predict the outcome of the general elections in the UK in the year 2010. He found that by aggregating and averaging the political opinions of subjects at the media level (Facebook, Google, Twitter, and YouTube), he was able to make predictions that were more accurate than traditional election polls. This indicates the usefulness of *the Wisdom of Crowds* in predicting outcomes in different fields.

Another important field in which the concept *the Wisdom of Crowds* has been applied in recent years is in betting markets. There is a vast amount of literature in which this concept is used as an attempt to better predict the outcomes of sports events. Brown and Reade (2019) looked into the use of *the Wisdom of Crowds* by analyzing predictions on sports events by amateur bettors. They found that selecting sections of this crowd that performed well in the past, does not significantly improve returns on future bets. However, when betting on sports outcomes using the majority vote of the crowd, average returns over 68,339 events turned out to be 1.32%. This indicates that averaging the predictions of crowds, often gives predictions of sport outcomes that are correct.

In recent years, many papers have tried to improve the predictions on the outcome of sports events by using different methods of aggregation of the individual estimates within groups. As mentioned before, selecting individuals who had good predictions in the past does not significantly improve returns on future bets. For this reason, it is interesting to investigate

other aggregation methods that could give better predictions. The betting markets are an industry in which a large amount of money is spent. These markets have been of high interest in *the Wisdom of Crowds* literature. International betting markets have been continuously growing over the years, and it is believed that global betting will have a market size greater than 94 billion dollar in the year 2024 (Lock, 2019). These growing betting markets globally give even more reason to investigate other aggregation methods to predict sports outcomes.

Concluding previous parts, aggregation of estimates of subjects within a group can give predictions that are often more accurate than any of the individual estimates. However, aggregating by averaging or taking the median do not seem to be robust methods. In some situations such as in sports betting they can give good predictions, but in other situations, aggregating estimates from subjects that have predicted well in the past could give better estimates. Using *the Wisdom of Crowds* concept to predict the outcomes of sports events can be very useful. In the light of the growing betting industry, the sport betting markets are the field of interest in this paper.

In the previously mentioned approaches, subjects were asked to give their own estimate of a certain value or outcome that had to be guessed. Prelec, Seung, and McCoy (2017) came up with a new approach to extract wisdom from a crowd. Previous statistical aggregation models used averages or median values of the estimates of all subjects within a group. Some of the researchers even included levels of confidence in own estimates. Their paper, however, has a different approach to extract the wisdom from the crowd. This new method, called the *Surprisingly Popular* method, offers a solution in situations in which the majority of a group believes in the wrong answer. In addition to asking subjects for a certain prediction, they also ask subjects what they believe is the most popular prediction among the other subjects and whether others will give the same answers or not. When an individual has more information than others and knows that others are likely to be wrong, both of these pieces of information can be expressed using the *Surprisingly Popular* method (Prelec et al., 2017).

As is shown in the previously mentioned paper, the *Surprisingly Popular* method can be a useful method to elicit wisdom from within a crowd. However, *the Wisdom of Crowds* is not applicable in situations in which a large share of the groups exhibits the same bias. In sports betting, betters tend to bet on their home team more often. This phenomenon is called the *Home bias.* Stanek (2017) gives possible reasons for the exhibition of the *Home bias* by subjects (for the Czech sports betting market). One of the possible explanations is the *Optimism bias* in which people´s desires influence their expectations. For example, people overestimating the

probability of good events and underestimating the probability of bad events. E.g. overestimating the probability that their home team would win. Thus, the estimated probabilities of home teams winning are biased upwards. Additionally, people are not willing to bet against their favourite team even though their team faces a stronger opponent. This leads to less diversification of bets and the situation that people are not willing to bet against or for certain teams. It is important to know whether the crowd is biased in a certain situation. When many individuals within a group are biased, predictions made using aggregations of all the estimates within the group are also likely to be biased.

By aggregating estimates of individuals within a group, predictions can be improved. As mentioned before, averaging or taking the median estimate of a group are not robust methods in giving correct predictions. The concept of *the Wisdom of Crowds* is not applicable in every situation, especially not in situations in which a large share of a crowd exhibits the same bias. A major field of interest in *the Wisdom of Crowds* literature is betting. Many papers have used aggregation methods in an aim to improve predictions of crowds. The *Surprisingly Popular* method showed to give promising results than other more basic aggregation methods. However, it is also proven that the *Home bias* plays an important role in betting behaviour. This gives reason to question whether the *Surprisingly Popular* method could improve predictions by controlling for this *Home bias.* For this reason, the following research question was created:

"Does controlling for the Home bias by using the Surprisingly Popular method improve the predictions of crowds on the outcome of sports events?".

To test this research question, three hypotheses have been formulated of which the first hypothesis is:

"Subjects mainly choose options that favor their favourite team and not options that are negative towards the performance of this favourite team".

Previous literature has shown the ability of the *Surprisingly Popular* method in correctly predicting the outcome by aggregating estimates within a group. Additionally, the existence of a *Home bias* in betting markets has been proven, for this reason it is believed that controlling for this bias will give better predictions in situations in which this bias is present. It is expected that subjects are more likely to agree with statements that are positive about the performance of their favourite team. The second hypothesis is the following:

"The Surprisingly Popular method gives more accurate predictions than averaging estimates of all subjects".

Prelec, Seung, and McCoy (2017) showed that the *Surprisingly Popular* method often gives good predictions even when a large share of the groups exhibits a bias, so it is believed that predictions using this model are better than the method of averaging estimates or taking the majority vote. Considering the fact that subjects are often biased in their judgements. The third and final hypothesis is:

"The Surprisingly Popular method gives more accurate predictions than confidence-weighted aggregation methods".

Another famous method of aggregation is using confidence-weighted predictions, as is done by Lee, Vi, and Danileiko, 2017. The confidence levels associated to agree and disagree answers are often similar for statements. It is expected that the *Surprisingly Popular* method will give better predictions than using these confidence-weighted predictions.

Past research has shown the predictive power of the *Surprisingly Popular* method in different fields. It has also been shown that the *Home bias* can play an important role in how people predict sport outcomes. The novelty of this research lies in the fact that it will investigate the use of the *Surprisingly Popular* method in predicting the outcomes of sports events, to control for this *Home bias.* No previous research has investigated this. Unfortunately, it was not possible to include actual bets on future sports events due to Covid-19. Therefore, the choice was made to use historical data instead.

The next section is the literature review, which will elaborate on how *the Surprisingly Popular* method has been applied in previous work. The subsequent section will go into the experimental design, especially how the data is gathered and analyzed. Thereafter, the findings of the statistical analysis of the data will be presented in the results section. Lastly, the paper ends with the discussion section which includes the conclusion, the implications of this research, and the future recommendations.

Literature review

The *Surprisingly Popular* method picks the answers that have more observed agreement than is expected. The idea behind this model is the following: When confronted with, for example, the following statement: ´Eindhoven is the capital of the Dutch province of North Brabant´, it is likely that many subjects agree with this statement. This is because Eindhoven is the largest city in the province. However, Den Bosch is actually the capital of the province of North Brabant. People who falsely believe Eindhoven is the capital of North Brabant are likely to believe that others agree with them. Subjects that have more information and know that Den Bosch is the capital, will likely expect others to falsely believe the capital is Eindhoven and rate the statement as true. When subjects falsely believe Eindhoven is the capital the expected agreement will likely be high. However, the observed agreement will be lower due to other people having more information and knowing it is actually Den Bosch. The surprisingly popular answer will be to disagree with the statement that Eindhoven is the capital of North Brabant because it has less observed agreement than expected agreement. When the statement is an easy one, the *Surprisingly Popular* method will be likely to still pick the correct answer due to a high observed agreement.

When some subjects have more information than others and the group is likely to exhibit a bias in predicting, the *Surprisingly Popular* method offers a solution. The model is less affected by the different forms of groupthink due to not simply picking answers that have the most support, but by using the additional information provided by the metacognitive follow-up question. This method combines the cognitive and metacognitive judgements, because subjects are asked to make an estimate of their own beliefs and think about the choices of other subjects. Furthermore, the model has no access to the correct answer, but it uses the difference in agreements to determine the *Surprisingly popular* answer (Lee, Danileiko, & Vi, 2018). In this paper, it uses the competence of a share of the subject pool in giving the right answer when rating whether statements based on sports events are true or not.

Often subjects exhibit the *Consensus bias*, in which they see their own prediction and judgements as common and appropriate. When having no information, it is likely that subjects believe that their answer is common and appropriate (Ross, Greene, & House, 1977). However, when a person with more information realizes that a large share of the group mistakenly rates a certain statement as true. This person will likely pick not true and believe that a small share of the subject group will agree with him.

The *Surprisingly Popular* method compares the expected proportion to the observed proportion. The latter is the percentage of subjects that give a particular answer. The former is the combined estimated percentages for all subjects. When people wrongfully think something is right, they often falsely believe the expected agreement is high. The decision made using the *Surprisingly Popular* method is to choose the answer that has a higher percentage of observed agreement than is expected. This answer is the surprisingly popular answer (Lee, Danileiko, & Vi, 2018). The *Surprisingly Popular* method has shown to give accurate predictions.

Lee, Danileiko, and Vi (2018) tested the ability of the *Surprisingly Popular* method in predicting sports events, namely the 2017 NBA (National Basketball Association) playoffs. This was the first time the algorithm was tested in a situation for which at the time was no right answer. Thus, making genuine predictions for these events. They used the *Surprisingly Popular* method to predict the matchups in the 2017 USA NBA playoffs. The predictions of this method were then compared to a confidence-weighted aggregation method, and the average estimate predictions.

All of these methods showed to be effective in predicting the correct outcomes. This was mostly due to the fact that the widely favored teams won all the matches (Lee, Vi, & Danileiko, 2017). Seeing that all these matches are won by the widely favored teams, the results of this research are not that informative. Lee, Danileiko, and Vi (2017) stated that in a situation in which a subset of the subjects has an insight into a surprise winner, the *Surprisingly Popular* method would capture that knowledge. However, they were uncertain whether and how often these subsets of subjects would exist.

The same authors tested the ability of the *Surprisingly Popular* method in predicting the outcomes of NFL games (National Football League). Previously, the *Surprisingly Popular* method showed to be performing well in situations in which the correct answer was already known. In their paper, they investigated the use of this method to predict outcomes of new events. They used the method to predict the outcomes of 256 NFL games in the season of 2017-2018. All the subjects that participated in this research stated their own ability to be ´extremely knowledgeable´ on this topic. They compared the *Surprisingly Popular* method to other aggregation methods such as averaging predictions and found that the *Surprisingly Popular* method outperformed other methods. These other methods were confidence-weighted estimates, averaged estimates, and an aggregation method based on past performance (Lee, Danileiko, & Vi, 2018).

Furthermore, they compared the performance of these methods to benchmark data.

More specifically, the predictions of a group of 94 experts, and a forecasting website which uses an algorithm that is similar to the ELO method used in chess. The *Surprisingly Popular* method seems to outperform most of these experts, the majority vote of these experts, and the predictions of the algorithm similar to the ELO method. Seeing that this research is focused on one season of the NFL, it can not be concluded that the *Surprisingly Popular* method outperforms the other methods in different settings. Lee, Danileiko, and Vi (2018) deemed this study to be a motivating demonstration of the *Surprisingly Popular* method.

In the research by Görzen and Laux (2019), an extensive comparison is made of different aggregation methods to extract the wisdom from crowds. In the experiment they conducted, different methods were compared on their ability to give the correct answers. The different aggregation methods that they compared are the following: the majority vote, the confidence-weighted predictions, confidence only prediction, average confidence prediction, and the *Surprisingly Popular* method. An interesting fact about their research is that they used a crowd through a commercial crowdsourcing website.

Subjects in their experiment were tasked to rate 35 general knowledge statements as true or false. They found that the majority prediction had the lowest percentage of correct predicted answers, and the average confidence prediction the highest. Furthermore, it was expected that the *Surprisingly Popular* method would have the best performance based on the previous literature. However, this was not the case in this experiment. A possible reason for the underperformance of the *Surprisingly Popular* method stated in this paper is the anonymity of the crowd that is used. The limitations of this research are that the used questions could have been too specific. Therefore, the results are not generalizable. Furthermore, since it was an online crowd, people could have looked up the correct answers (Görzen & Laux, 2019).

McCoy and Prelec (2017) proposed another model to aggregate the answers of subjects while incorporating the judgements of peers. In contrast to the *Surprisingly Popular* method, their model focuses on aggregating the answers of subjects answering both single answers as multiple-choice questions. Similar to the *Surprisingly Popular* method, their model does not assume that all subjects have access to the same information and therefore it does not assume that the majority vote is always correct. Subjects in this experiment are modelled as receiving one single signal on the actual state of the world and they use this signal to give their own prediction and a prediction of what they believe others will predict. McCoy and Prelec (2017) call the model they propose in their research the *Possible World*

Model, because it does not only take into account how people vote, but it also considers the vote distribution in all possible worlds (McCoy & Prelec, 2017).

In their paper, the performance of their model based on the single question answers is compared to the majority vote, the *Surprisingly Popular* method, and linear and logarithmic prediction pools. Additionally, they compare the performance on multiple-choice questions to some other multiple-choice questions models that are of less interest. The models are evaluated using data from seven different studies, in which participants were tasked to give their own prediction, and a prediction of what they believed what others would do. The precision in terms of choosing the correct answer out of two possibilities on separate questions is not significantly different for the *Possible World Model*, and the *Surprisingly Popular* method. However, the *Possible World Model* does not seem to perform well when the answers to questions maintain a consistent ordering.

Concluding this section, the performance of the *Surprisingly Popular* method has been assessed in several studies in recent years. In most of these studies, this method seems to outperform other aggregation methods. However, it is not clear in which settings the *Surprisingly Popular* method performs the best. Furthermore, the results of previous work on applying the *Surprisingly Popular* method on sports events is not generalizable. It is not certain whether using this method would yield improved predictions by controlling for a possible *Home bias* in predicting behaviour. It is interesting to see that the model is applicable in many different scenarios.

Methodology

An experimental design is used to test the hypotheses and answer the research question. A survey is conducted among subjects, questioning them on historical football statistics of football clubs in the Dutch Football league the *Eredivisie*. This survey is distributed through online channels to reach as many subjects as possible. Furthermore, this data is analyzed using several statistical methods. The intention of this paper was to use actual bets on future events. However, this was not possible due to the uncertainty caused by Covid-19 and the resulting absence of sports events in spring 2020. Considering the circumstances, it was chosen to use historical data instead of fixtures.

Data collection

The survey contained historical data of Dutch football teams and their matches in the highest Dutch football league the *Eredivisie*. These are statistics which subjects were asked to rate as true or false (binary). Considering, that it was likely that the majority of subjects in this research would support a Dutch team, the focus of these statistics was on Dutch football teams. Hence, it was more probable that participants were asked to rate the trueness of statements that were based on the results of their favourite team, and therefore possibly exhibit the *Home bias* in their predicting behaviour. The sports statistics used in this experiment are based on overviews of the results of football matches between clubs in the Dutch Football league the *Eredivisie* over the last 10 seasons, from season 2009/2010 up until 2019/2020, including all finished matches of the season 2019/2020 at the date May the 1st, 2020.

Individuals were asked to make these judgements on their own without looking up the answers on the internet. Furthermore, they received no information on the judgements of other subjects in this survey. This was done to make sure that subjects are not influenced by the other subjects. The social influence could have an impact on the statistical aggregate and could even destroy the wisdom within a crowd due to members of the crowd revising their own estimates when confronted with answers of others (Mason, Conrey, & Smith, 2007). Lorenz, Rauhut, Schweitzer, and Helbing (2011) showed that social influence can have multiple effects that undermine the wisdom of the crowds. Their main finding is that social influence is able to trigger the convergence of estimates. Thus, reducing the diversity of the group, while not improving the accuracy of the predictions within this group. The heterogeneity of a group often creates a more accurate prediction than predictions by a group ruled by social influence. In the interest of reducing the social influence on the subjects and avoiding the negative effects on the accuracy of the predictions made by the group, subjects are requested to fill in the survey on their own. Afterwards, these judgements were aggregated and put to the test using a multiple of algorithms.

The focus of these historical data questions is on the top five best male football clubs in the Netherlands. This ranking is based on the difference in the ELO ratings of these clubs. The following Dutch clubs were used: Ajax, AZ, FC Utrecht, Feyenoord, and PSV. The club with the highest ELO rating in the Netherlands is Ajax with 1767 ELO points, and the club with the lowest number of ELO points included in this research is FC Utrecht with 1516 ELO points. All these football clubs are competing in the same league. For this reason, the statistics used are of matches played within this league. Other matches of these teams, such as friendly

matches, are not included in this research. The performance of these teams in these matches could be of less importance. Therefore, this could create a bias in the statistics, due to clubs not always performing at their best in certain types of matches by not putting in full effort.

The ELO rating is a rating system that rates clubs or players based on their past performance. Therefore, this rating is not influenced by human judgements (Lee, Danileiko, & Vi, 2018). It was chosen to use the ELO rating to approximate the top five best football clubs for the reason that this measure is based purely on past performance and relative skill levels. The difference in ELO points between two clubs predicts the outcome of a match played between those clubs. The table below (Table 1), shows the ELO ratings of the clubs that are used in this research. All these football clubs have been in the *Eredivisie* continuously for at least the last 10 seasons.

Table 1

Top five Dutch football clubs ELO

Club	ELO points
Ajax Amsterdam	1767
PSV Eindhoven	1643
AZ Alkmaar	1625
Feyenoord Rotterdam	1610
FC Utrecht	1516

Historical data statements

All the data used for the statements in the survey is retrieved from the website http://www.eredivisiestats.nl/ (Hulsen, 2020). An overview of the results of the matches between the top five clubs over the last ten seasons in percentages is given in Table 2 on the next page. The subjects participating in the survey received a statement similar to the following: *"Feyenoord has won more than x% of their matches against Ajax in the last 10 seasons*

(2010/2011-2019/2020, including all matches finished up until the 1st of May, 2020)". When confronted with this statement, subjects were asked to rate the statement as true or false. The possible answers they were able to give as a response to these statements were "agree" or "disagree". The subjects were asked to rate a total of ten questions similar to the one mentioned above, differentiating the clubs, the percentages and whether the club has lost or won. An overview of all the statements used in the survey can be found in appendix A.

Table 2

Historical data matches top five

	Number of games played						
	Won		Draw		Lost		Total
Club	Nw	%	Nd	%	Nl	%	Nt
Ajax - *AZ*	10	50.0%	4	20.0%	6	30.0%	20
Ajax - *Feyenoord*	12	63.2%	5	26.3%	2	10.5%	19
Ajax - *PSV*	9	45.0%	5	25.0%	6	30.0%	20
AZ - *PSV*	4	21.0%	0	0.0%	15	79.0%	19
AZ - *Feyenoord*	5	26.3%	4	21.1%	10	52.6%	19
FC Utrecht - *Ajax*	6	31.6%	6	31.6%	7	36.8%	19
FC Utrecht - *AZ*	8	42.1%	3	15.8%	8	42.1%	19
FC Utrecht - *PSV*	2	10.5%	2	10.5%	15	79.0%	19
Feyenoord - *FC Utrecht*	13	65.0%	6	30.0%	1	5.0%	20
Feyenoord - *PSV*	10	50.0%	2	10.0%	8	40.0%	20

Note. How many games **A** has won/played draw/lost against *B*

Table 2 on the previous page**,** gives an overview of the results of the matches between the five clubs. Due to the outbreak of COVID-19 in early 2020, the Eredivisie was discontinued and for this reason, not all clubs played against each other twice in the season 2019/2020. Therefore, the total number of matches played between certain clubs differs from the number of matches played between other clubs. The clubs that were able to play both their matches against another club in the season 2019/2020, played a total of 20 matches against this club over the last ten seasons. However, the clubs that were not able to play both matches, played one match less in the same timeframe. As can be seen in the table, the highest percentage of wins is 79% achieved by PSV against AZ, and against FC Utrecht. Furthermore, the highest number of draws is 31.6% for matches between Ajax and FC Utrecht.

The Surprisingly Popular method

Subjects were also questioned on what they think other subjects will predict to be the outcome of these events, as is done in the paper by Prelec, Seung, and McCoy (2017). This was done in light of the *Surprisingly Popular* method, to measure the expected and observed proportions. Therefore, subjects were asked to estimate the percentage of people that predict the same outcome as they do for each statement (expected agreement). The observed proportion is the actual proportion of people that agree with a statement and the expected proportion is the predicted proportion that gives a particular answer as estimated by all the subjects. The question used to measure the expected proportion of subjects that agree is the following: *"What percentage of other subjects do you think has chosen this statement to be true? "*. The observed proportion is measured by measuring the number of subjects that rate a statement as true or false. Each question asking subjects whether a statement is true or not is accompanied by one of these metacognitive questions.

Confidence in estimates

The survey continued with questions measuring the level of confidence in the judgements of subjects on the ten statements. This was done by asking subjects how certain they were in their judgement on whether the statement on the historical data of the football clubs in percent was true or not. When a subject states to be completely uncertain in predicting the outcome, there is still a 50% chance they pick the right outcome. For this reason, the answers they were able to choose from were percentages from 50% upwards. Furthermore, questions were added to measure preferences/attitudes towards all the different teams that are included in this research. Table 3 on the next page, shows the possible confidence levels that subjects were

able to choose from. Each question asking whether a statement is true or not is followed by a question measuring the confidence of the subjects in their answer on that question.

Table 3

Rate the level of confidence in own answer on the previous question

Answers confidence measure	
Answer	Corresponding %
Totally uncertain	50%
Low Confidence level	60%
Moderately Confident	70%
High Confidence level	80%
Very high Confidence level	90%
Certain	100%

Demographic variables

The survey ended with questions measuring demographic variables. These demographic variables were measured to be used as control variables in the statistical analysis of the data and to give descriptive of the subjects in the experiment. The demographic variables that were measured are the following: age, gender, level of education, and nationality. It is likely that these questions cost less mental effort, and for this reason the survey ended with these questions. Furthermore, the subjective level of knowledge of the subjects on football and the league *Eredivisie*, and their involvement with the sport is measured. Their subjective knowledge is measured by asking them how much knowledge they have about football and on the league. In addition to this, the subjects were asked whether they play football or have played it in the past. Also, to determine whether subjects support one of the clubs in the research they were asked to pick their favourite team within the *Eredivisie*. These variables were measured to be possibly

used as control variables in the analysis of the data and to have a closer look at the possible *Home bias*. The data extracted from Qualtrics is statistically analyzed using other programs. Furthermore, the questions measuring the subjective involvement of subjects with Football and the Dutch football league are used to compare the performance of the algorithms on different subgroups within the total group of subjects. Table 4 below, gives an overview of the demographic and the football-specific questions asked, and which variables they measured.

Table 4

Control questions and measured variables

Question	Variable measured
What is your age?	Age in years
What is your gender?	Gender subject
What is the highest level of education you completed?	Level of education
What is your Nationality?	Nationality
How much knowledge do you have about Football?	Knowledge game
How much knowledge do you have on the *Eredivisie*?	Knowledge league
Do you play football or have you ever played football?	Experience
What is your favourite football team in the Eredivisie?	Favourite team
Do you actively follow football games in the Eredivisie?	Involvement league

Data analysis

The data gathered through the survey is statistically analyzed using multiple methods. In addition to the *Surprisingly Popular* method, the choices of the subjects are analyzed using the following other algorithms: confidence-weighted prediction method, most-confident prediction, and the majority prediction. Using these algorithms allows to compare the predictive

power and accuracy of these models with the *Surprisingly Popular* method. These algorithms will be explained more thoroughly in another section of this paper.

The subjects were asked to rate the trueness of certain statements on historical data of football matches between certain clubs. After rating whether statements were true, the subjects were asked to estimate the percent of other subjects that choose the same answer. The answers to these two questions will be used for the *Surprisingly Popular* method. The measured observed agreement will be compared to the expected agreement as given by the subjects. The performance of the *Surprisingly Popular* method will be compared to other aggregation models.

Home bias

People are likely to predominantly bet on the home team and not against it (Staněk, 2017). When rating the trueness of statements regarding their favourite team, subjects are more likely to rate statements that are positive towards the performance of their team as true and statements that are negative towards their team as false. Therefore, subjects were asked to state which team in the *Eredivisie* is their favourite team. This information is used to determine whether subjects were more likely to give answers that are in favor of their favourite team. This is included to test the first hypothesis: *"Subjects mainly choose options that favor their favourite team and not options that are negative towards the performance of this favourite team"*.

Majority vote

The majority vote or majority prediction is the aggregation model that picks the answers or choices which received the most support by the subjects. In this particular research, subjects were asked to rate statements as true or as false. This makes these statements binary, seeing that the only possible answers are true or false. The majority vote will be picked by comparing the total number of supporters for each of the possible answers. The outcome that has the most support is the majority prediction or majority vote. This aggregation method can in some situations give good estimates of what the right answer would be. However, in some cases when there is groupthink, this method can give false and biased predictions. For this reason, this method will be compared to the other aggregation models. The majority prediction will be estimated by comparing the total number of answers for both answers, picking the largest number out of the two. The ability of this model to predict the correct outcome is compared to the *Surprisingly Popular* method. This method is used to test the second hypothesis which is stated as follows: *"The Surprisingly Popular method gives more accurate predictions than averaging estimates of all subjects"*.

Confidence weighted predictions

Subjects were asked to state their confidence in their choices for each of the statements. The possible choice outcome for rating the statements is binary. Seeing that a guess would amount to a ½ chance of guessing the right answer, it is not possible for a subject to state confidence levels below 50%. Different methods can be applied with these gathered confidence levels. The first one that will be compared to the *Surprisingly Popular* method, is picking the answers for which subjects on average state to have the most confidence. This algorithm is called the confidence-weighted prediction method. Calculating these confidence -weighted predictions is done by taking the overall average of confidence per possible answer for the whole group of subjects, and picking the answer for which subjects rated to be more confident. The other confidence-weighted method that is used in this paper, is picking the answer that has the most support by subjects who stated to be completely certain in their answer. When asked to give the level of confidence in their answer, the highest possible level of confidence subjects are able to choose is being certain (100%). The reason that more people who state to be completely certain in their answer choose a particular answer might be because they have more information than others. The ability of these algorithms to give the correct answers will be compared to the *Surprisingly Popular* method. These algorithms which are named the most-confident prediction, and the confidence-weighted prediction will be used to test the third hypothesis which says that the *Surprisingly Popular* method gives more accurate predictions than confidence-weighted predictions will give.

Models used

Several models will be used in an attempt to test the three hypotheses and give an answer to the research question. Also, a regression model will be used to have a closer look at the relationship between some of the demographic variables and the ability to correctly score the statements as true or not.

Firstly, we will have a look at hypothesis: h1:

> *"Subjects mainly choose options which favor their favourite team and not options which are negative towards the performance of this favourite team".*

A comparison of the predictions of subjects on statements containing their home team and whether their predictions are for or against their team will be made. This comparison should give some evidence for whether people more often choose good outcomes for their home team

than bad outcomes. Considering that all five clubs in this experiment are compared to other top five clubs, it would not make sense that for all clubs the majority of supporters predicts more positive outcome without a *Home bias* existing. This comparison will not give a decisive answer on whether the subjects in this experiment exhibit a *Home bias*. However, it will give some information on how these subjects make decisions.

A binomial test (with proportion 0.5 under the null hypothesis) will be performed to compare the difference in judgements of subjects in relation to their favourite team. Looking into whether people significantly choose options that favour the performance of their team. In addition to this, multiple paired sample t-tests are performed to test whether there is a significant difference in performance for when people rate statements that include their favourite team against the performance on the other statements. These paired sample t-tests will be performed on the groups of subjects that stated their favourite team to be one of the five clubs used in this paper. In contrast to the binomial tests, these paired sample t-tests could tell us whether these subjects exhibit a *Home bias*, because they actually show if subjects perform differently when rating statements that include their favourite team.

Secondly, the four algorithms will be compared on accuracy by showing their basic precision in a graph. The prediction accuracy of the four algorithms on the total subject group will be shown side by side. As mentioned before, the four algorithms of interest are the *Surprisingly Popular* method, the majority vote, the confidence-weighted prediction, and the most-confident Prediction. This graph will only show the ability of the algorithms to give correct answers on the ten statements in percent. This will be done in a first investigation into the second hypothesis: h2: *"The Surprisingly Popular method gives more accurate predictions than averaging estimates of all subjects"*, and the third hypothesis: h3: *"The Surprisingly Popular Method gives more accurate predictions than confidence-weighted aggregation methods"*.

Classification accuracy

For the reason that percentual agreement could be high by chance, the classification accuracy is further assessed by categorical correlation coefficients. The measures that will be used to assess the categorical correlation are the F1-score, Cohen's kappa, and Matthews Correlation Coefficient. The precision of the algorithms mentioned in the previous part could be partly due to chance. Using these three measures could give some more information about the accuracy of the four algorithms. The assessment of the classification accuracy using these particular tests is inspired by the paper by the paper by Prelec, Seung, and McCoy (2017).

F1-score

Firstly, the measure F1-score will be used to assess the accuracy of the algorithms. The F1-score is a measure of the accuracy of binary classification. When computing a score, this test takes both recall and precision into consideration. In which precision is the number of correct positive results divided by all positive results given by the classifier. The recall is the number of positive results that were correctly classified, divided by the total number of results that should have been classified as being positive. The score is not influenced by the number of correctly classified negative results (Chicco & Jurman, 2020). When the score takes a value of 1, precision and recall are considered perfect, a value of 0 implies that the classification performed in the worst way possible. The formula of the F1-score is given below.

Formula F1-score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Matthews Correlation Coefficient

Secondly, the measure Matthews Correlation Coefficient (MCC) will be used to assess the accuracy of the binary classification of the algorithms. As opposed to the F1-score, the MCC takes into account both true positives/negatives, and also the false ones. The F1-score only takes into account the correct positive results. The scores given by this measure can take values in the range of [-1,1], in which 1 indicates a perfect prediction by the classification model. A value of 0 implies that the prediction made by the classifier is the same as a random prediction, and a value of -1 represents a complete disagreement between prediction and reality. The formula of the MCC is given below.

Formula Matthews Correlation Coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Cohen's kappa

Thirdly, Cohen's Kappa Coefficient will be used to assess the accuracy of the binary classifier. This measure takes into account that correct predictions could have been due to chance. The score given by this measure can take a score in the range of [-1,1], in which 0 would imply that the predictions are not better than a random prediction would give. The baseline of Cohen's kappa is the percentage of agreement due to random allocation. Cohen's

kappa coefficient was initially used to measure the degree of agreement between observers of the same event (Cohen, 1960). However, it can also be used to compare the performance of one observer to reality. The Cohen's Kappa coefficient is measured using table 5 and the formulas that can be found below. A Cohen´s kappa of 0-0.20 indicates a slight agreement between model and reality, a score of 0.20-0.40 a fair agreement, 0.41- 0.60 a moderate agreement, 0.61- 0.80 a substantial agreement, and 0.81-1 an agreement that is almost perfect or even perfect.

Table 5

Cohen's kappa

		B	
		Yes	No
A	Yes	A	B
	No	C	D

Formulas Cohen´s kappa:

$$\kappa = \frac{Po-Pe}{1-Pe} = 1 - \frac{1-Po}{1-Pe}$$

$$Po = \frac{a\sim + d}{a+b+c+d}$$

$$Pe = Pyes + Pno$$

$$Pyes = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d}$$

$$Pno = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$$

McNemar Test

In an attempt to have a closer look at the accuracy of the algorithms, these algorithms are compared using the McNemar test. The McNemar test will be used to compare the performance of the *Surprisingly Popular* method separately with the other three algorithms. This test is useful when confronted with repeated measures of two related groups, which is the case in this experiment. Furthermore, the dependent variable is a dichotomous variable that can take the values correct or incorrect. The McNemar test is used to test whether there is a difference in these dichotomous dependent variables between two models (Trajman & Luiz, 2008).

Using the current design, the number of total values that can be used in the McNemar test is fairly low with a value for each of the four algorithms on each of the statements. To increase the number of values for the McNemar test, the total sample of subjects has been randomly divided into four equally sized groups. This is done to increase the predictive power of these tests. As mentioned before, the subjects are questioned on a total of ten statements, therefore this would leave us with a maximum of ten values per used algorithm. By splitting the group into four new groups allows to increase the number of statements and thus reach the desired number of values. By creating four groups this leaves us with approximately 40 values per algorithm. The test is applied to a 2x2 contingency table to check whether the column and row marginal frequencies are equal. An example of this sort of contingency table used for the McNemar test is shown below in table 6. The McNemar test is a non-parametric test. Thus, it makes no assumptions about the parameters of the population.

Table 6

Contingency table McNemar test

	Test 2 True	Test 2 False	Row total
Test 1 True	a	b	a+b
Test 1 False	c	d	c+d
Column total	a+c	b+d	N

Kruskal-Wallis test

As mentioned above, the total group of subjects has been randomly divided into four smaller groups. To test whether these newly created groups are on average similar, a Kruskal-Wallis test will be performed on the variables age, gender, level of education, and the subjective knowledge of the subjects on the *Eredivisie*. The Kruskal-Wallis test is a nonparametric test. Thus, no assumptions are made about the parameters of the population. The advantage of this test is that the test is general and easy to calculate. However, nonparametric tests tend to waste information, and are less sensitive than other tests. The outcome of this test will show if there is a significant difference in these variables between the four created groups or not.

Explorative analysis

To have a look into the different demographic variables and the ability of subjects to correctly predict whether statements were true, a regression analysis has been performed including some of the control variables asked through the demographic questions in the survey. This is done to investigate whether people with certain characteristics are more likely to give correct answers than others. The dependent variable in this regression model is the computed variable representing the number of correct judgements of the ten statements. The independent variables used in this regression model are age, gender, level and education, and the subjective knowledge on the *Eredivisie* of the subjects. The regression formula is as follows:

$$Score = \alpha + \beta_i 1 \cdot Knowledge_i + \beta_i 2 \cdot Age_i + \beta_i 3 \cdot Education_i + \beta_i 4 \cdot Gender_i + \varepsilon_i$$

When independent variables in a regression model are correlated there is multicollinearity. However, independent variables should not be correlated with other independent variables. Multicollinearity can cause problems when fitting the model and interpreting the results (Alin, 2010). A regression model is used to examine the relationship between the dependent variable and the independent variables of interest and shows how the dependent variable would change when one of the independent variables changes with one unit, keeping all other factors constant. When there is correlation between some of the independent variables, a one unit change in these variables could also lead to changes in the values of other independent variables and through these other independent variables on the dependent variable. When the correlation between independent variables becomes larger, this makes it more difficult to measure the relationship between the independent variables and the dependent variable.

The involvement with (Dutch) football variables used in this research are likely to be highly correlated with each other. For example, the variable measuring how often people watch football games and the variable measuring how often people watch *Eredivisie* games could be perfectly correlated in some situations. Therefore, some of these variables are left out of the exploratory regression model. It is chosen to pick one variable out the bunch of variables that are much alike, which is likely to be the best predictor of a higher ability to correctly predict correct outcomes on the statements. This variable is the subjective knowledge of subjects on the Eredivisie. Additionally, the control variables age, gender, level of education, and nationality are added. Seeing that these variables could account for some of the accuracy in predictions. It is not likely that these variables are correlated.

As mentioned before, the dependent variable in this regression model is score, which is a computed variable from the answers on whether the ten statements are true or not. For each correct answer this score increases with one point, the highest possible score is therefore ten points. There were two subjects in the sample with a score of nine which is the highest within the total sample, the lowest score is for a person with zero points.

Data

Descriptive statistics

From the total of 249 subjects who started the survey, 158 people finished it. All subjects that did not completely finish the survey were removed from the sample. When asked about their age, one subject stated their age to be below 18 years old. 31 subjects stated their age to be between 18 and 21 years old, 92 subjects stated their age to be 22-25 years old, and 34 stated their age to be above 25 years old. The largest share of subjects lives in Rotterdam with 79 subjects, and the second largest group consists of eight subjects who live in Utrecht. Subjects reported living in a total of 43 different cities. Also, the largest share of subjects stated Rotterdam to be their home city with 49 people. Subjects named a total of 53 different cities, which they rated as being the city they feel at home in. Furthermore, the subjects were asked to name their favourite Dutch team in the Eredivisie of which an overview is given on the next page in table 7**.**

Table 7

Favourite Dutch football team in the Eredivisie

Club	Frequency	Percent	Cumulative Percent
ADO Den Haag	1	0.6	0.6
Ajax Amsterdam	48	30.4	31
AZ Alkmaar	2	1.3	32.3
FC Emmen	1	0.6	32.9
Feyenoord Rotterdam	66	41.8	74.7
FC Groningen	1	0.6	75.3
PEC Zwolle	2	1.3	76.6
PSV Eindhoven	14	8.9	85.4
Sparta Rotterdam	4	2.5	88
FC Twente	2	1.3	89.2
FC Utrecht	11	7	96.2
Vitesse Arnhem	2	1.3	97.5
Willem II Tilburg	4	2.5	100
Total	158	100	

Nearly all subjects are of Dutch nationality with 151 subjects, six other subjects stated to being Non-Dutch but living in Europe, and only one person has a non-European nationality. The largest share of the subject pool is male with 111 subjects, followed by 46 subjects being female and one person stating their gender to be ´other´. The subjects were asked to state their highest level of education completed, table 8 in appendix A gives an overview of the highest level of education completed for all subjects. The largest group share of subjects stated the highest level of education completed to be a bachelor's degree with 55.1%, followed by people who only finished their High school with 28.5%.

Results

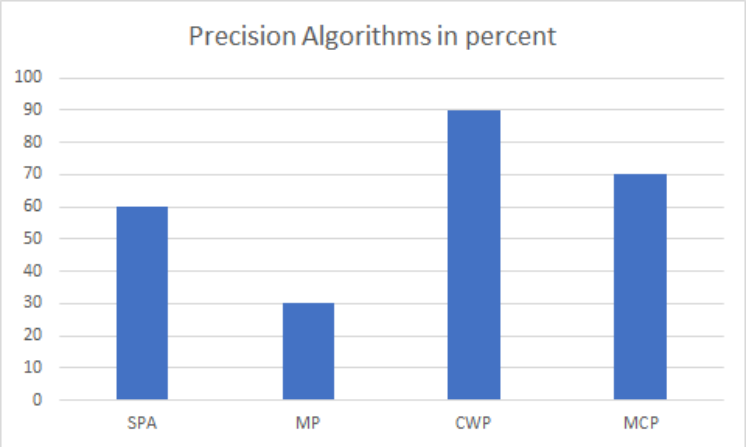
To test whether subjects were more likely to rate statements that are positive towards the performance of their favourite team as true and statements that are negative towards their team as false, multiple binomial tests have been performed. The test proportion used in these binomial tests is 0.5. Due to the small number of subjects (2) who stated AZ Alkmaar to be their favourite team, this club is left out of the binomial tests. At the significance level of 0.05, the clubs Ajax, PSV, and Feyenoord significantly deviate from the test proportion of 0.5, supporters of the clubs Ajax and Feyenoord more often rate statements that are positive towards the performance of their team as true and statements that are negative towards their team as false. However, supporters of PSV more often rate statements that are negative towards the performance of PSV as true. The corresponding P-values for these clubs are 0.000 for both Ajax and Feyenoord, and 0.002 for PSV. The output of these binomial tests can be found in Appendix A.

Following the binomial tests, a number of paired sample t-tests have been performed to look into the difference in performance for when subjects rated statements based on the performance of their favourite team, against their performance on the other statements, which not included their favourite team. It was chosen to perform these tests separately on the different clubs supporting groups of subjects.

For the supporters of the clubs Ajax and Feyenoord there was a significant difference in the mean of performance on statements including their favourite team against statements that did not, at the significance level of 1% with corresponding p-values of 0.000 for both Ajax and Feyenoord. There is also a statistical difference in means between these performances for supporters of PSV at the significance level of 10% with a p-value of 0.077. No significant difference in means are found for supporters of FC Utrecht at a significance level of 10%. Supporters of Ajax and FC Utrecht seem to have a higher performance on statements that did not include their favourite team. In contrast, supporters of the clubs PSV and Feyenoord seem to perform better on statements that include their favourite team. The output of these paired sample t-tests can be found in Appendix A.

Figure 1

Precision of the four methods



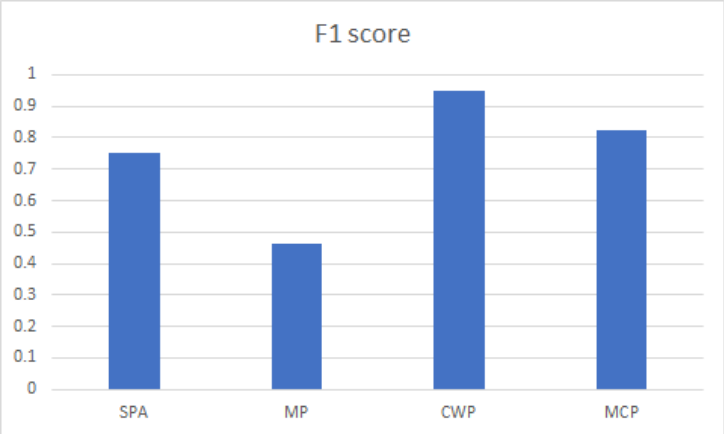
The precision of the four algorithms in predicting the correct outcome on whether the statements based on historical data were true or not is shown above in Figure 1. The figure shows that the confidence-weighted prediction method has the highest accuracy in correctly predicting the outcome by aggregating the estimates of all subjects. The least accurate algorithm is the majority prediction, which surprisingly predicted the correct outcome for only three out of the ten statements. The *Surprisingly Popular* method showed to be correct on six out of the ten statements. Lastly, the most-confident prediction method correctly predicted the outcome for seven out of the ten statements.

Following this, the classification accuracy is assessed using categorical coefficients. This was done for the reason that the accuracy of the algorithms could have been caused by chance. The first measure that has been applied to the predictions of the algorithms is the F1-score, this measure takes into account both the precision and the recall. This measure has been applied to the four algorithms on the total group of subjects. A F1-score of one is the highest a F1-score can reach, a score of one implies that the model is perfect in prediction and recall. In which recall is the number of correctly guessed positive results divided by all of the values that should have been positive, and precision is the number of correct answers divided by the total number of answers.

As can be seen in Figure 2 on the next page, the confidence-weighted prediction method has the highest F1-score with a value of 0.947, second is the most-confident prediction method with a F1-score of 0.824, followed by the *Surprisingly Popular* method with a F1 score of 0.75, and lastly the majority prediction method with a score of 0.461.

Figure 2

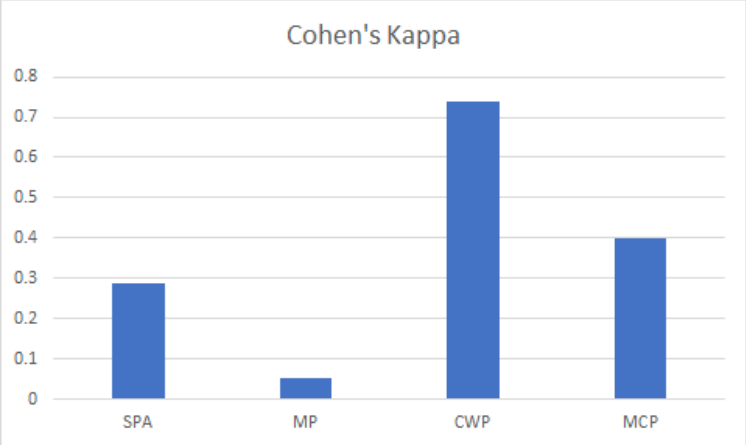
F1 scores of the four methods



Subsequently, the Cohen's kappa measure has been applied to these same algorithms, as a second assessor of the accuracy of these algorithms. As can be seen in figure 3 below, the confidence-weighted prediction method has the highest Cohen´s kappa with a score of 0.737, implying that there is a substantial level of agreement between the algorithm and reality. The second highest Cohen´s kappa is a score of 0.400 for the most-confident prediction method, which suggests a fair agreement between algorithm and reality. The *Surprisingly Popular* method has a kappa of 0.286, which also hints at a fair agreement. Lastly, the smallest Cohen´s kappa is a kappa of 0.054 for the majority prediction, which indicates that there is slim to none agreement between the algorithm and reality. When comparing the Cohen´s kappa of the different methods, it seems that the confidence-weighted prediction method give the most accurate predictions.

Figure 3

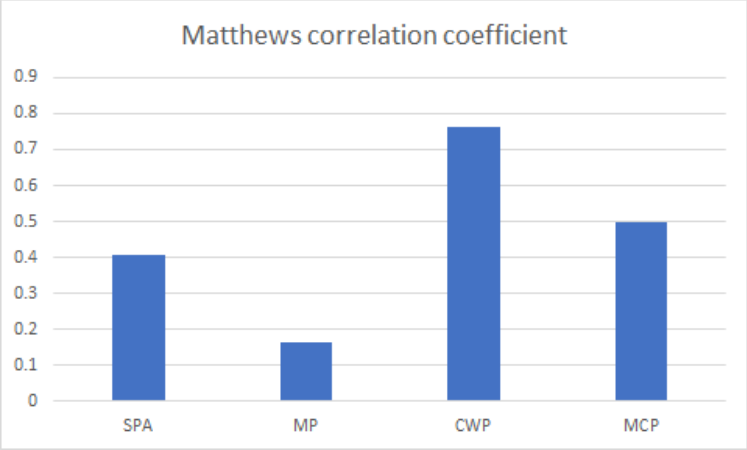
Cohen´s kappa of the four methods



Lastly, the Matthews Correlation Coefficient measure has been used to assess the accuracy of the algorithms in predicting the correct outcome. The highest coefficient found is for the confidence-weighted prediction method with a score of 0.764, indicating a very strong positive relationship between the predictions of the algorithm and the actual values, followed by the most-confident prediction method with a score of 0.500 indicating a strong relationship. The *Surprisingly popular* method has a score of 0.408, which also indicates a strong relationship. Lastly, the majority prediction has a score of 0.166 implying that there is no or a negligible relationship between the predictions of this method and the actual values. When comparing the method based on the Matthews Correlation Coefficient, the confidence-weighted prediction method again seems to give the most accurate predictions.

Figure 4

Matthews Correlation Coefficients algorithms



Kruskal-Wallis test

To test whether the four randomly created groups are actually random, the Kruskal-Wallis test has been used. The variables tested using this test are the following: nationality, gender, age, knowledge on the Eredivisie, and level of education. The output of the test can be seen in Table 9 on the following page. The p-values for these variables are as follows: 0.787 (gender), 0.542 (age), 0.622 (knowledge on Eredivisie), and 0.678 (level of education). The output of the Kruskal-Wallis test is not significant for any of the variables at a significance level of 10%. This implies that the four groups are not statistically different on the variables gender, age, knowledge on the *Eredivisie*, and level of education using a significance level of 10%. Thus, there is no evidence that the groups are not correctly randomly assigned on a significance level of 10%, based on these variables.

Table 9

Test Statistics of the Kruskal-Wallis Test

	Nationality	Gender	Age	Knowledge Eredivisie	Level of Education
Kruskal-Wallis H	1.495	1.06	2.151	1.768	1.519
df	3	3	3	3	3
asymp. Sign.	683	787	542	622	678

Note. Grouping variable: random group assignment.

McNemar Test

By splitting the total group of subjects into four new groups, new values were created. Some for which the different algorithms did not give a conclusive answer. For example, the majority vote was not applicable in some of the cases, because there was no majority prediction for each of the statements. Instead, for some of the groups the votes on some of the statements were equally divided among the two possible answers. Therefore, these values are left out of the test. For two of the statements, it was not possible to use the majority vote. Furthermore, seeing that there were a large number of statements for which there were no fully confident people. This aggregation method is left out of the McNemar test analysis part. Thus, the prediction accuracy of the *Surprisingly Popular* method is compared to the accuracy of the confidence-weighted method, and the majority prediction. Additionally, the majority prediction is compared to the confidence-weighted prediction.

The output of the McNemar tests (Appendix A) show that there is significant difference in performance of the majority prediction method when compared to the confidence-weighted model at a significance level of 1% with a p-value of 0.007. The confidence-weighted method is significantly better in correctly predicting the outcome than the majority prediction method. Furthermore, the confidence-weighted method is significantly different from the *Surprisingly Popular* method at a significance level of 10% with a p-value of 0.077. The confidence-weighted method is significantly better in correctly predicting the outcome than the *Surprisingly Popular* method. Lastly, the majority prediction method does not significantly differ from the *Surprisingly Popular* method in performance at the significance level of 10% with a p-value of 0.250.

Exploratory analytics

To examine the possible correlation between the different demographic variables and the ability of subjects to correctly predict whether statements were true, a regression analysis has been performed. The subjective measure of subject´s knowledge on the *Eredivisie* has also been included in this model. The dependent variable in this regression model is the computed variable *score* measuring the number of correctly judged statements. Table 10 below, shows the output of this regression model.

Table 10

Explanatory regression model output

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
Club	B	Std. Error	Beta	T	Sig.
(Constant)	3.27	0.98		3.34	.001
Knowledge Eredivisie	0.03	0.01	.46	5.31	.000
Gender	-0.25	0.35	-.06	-0.73	.467
Nationality	0.54	0.60	.07	0.91	.367
Age	0.09	0.23	.03	0.38	.706
Level education	-0.16	0.19	-.07	-0.83	.407

The variable knowledge on the Eredivisie has a statistically significant effect on the score of the subjects. At the significance level of 1 % with a p-value of 0.000. An increase of one point on knowledge on the *Eredivisie* increases the score of subject with 0.03 points, ceteris paribus. The variables age, gender, nationality, and level of education do not have a statistically significant effect on the dependent variable score at the significance level of 10%. Thus, the regression model shows that of these variables only the variable knowledge on *Eredivisie* has a significant effect on the score of a subject.

Discussion

In this paper, an attempt was made to improve the predictions of crowds on sports outcomes by using the *Surprisingly Popular* method, and therefore controlling for the *Home bias*. This is tested through an experimental design using a survey in which subjects were asked to rate whether statements based on historical sports data were true or not. A variety of models have been applied to test the hypotheses and seek an answer on the research question: *"Does controlling for the Home bias by using the Surprisingly Popular method improve the predictions of crowds on sport outcomes?".*

The results show that there is a potential *Home bias* in the predicting behaviour of subjects. The binomial tests showed that subjects who stated that they support Feyenoord or Ajax, chose outcomes that are positive towards their favourite team significantly more than statements which are negative towards the performance of their favourite team. Supporters of the club PSV chose options that were negative towards the performance of their team more often. However, this does not give evidence to prove the existence of a *Home bias*. This only gives a first insight into the predicting behaviour of the subjects. Therefore, paired sample t-tests have been performed to compare the performance of subjects on statements about their favourite team against statements that did not include their favourite team. These paired sample t-tests showed that there is a significant difference in performance between these two types of statements for some of the subject groups. Namely for supporters of the clubs PSV, Feyenoord, Ajax, and FC Utrecht. Supports of the clubs Ajax and FC Utrecht perform better on statements that did not include their favourite club, and supporters of PSV and Feyenoord performed better on statements that included their favourite club. These findings give reason to not reject the first hypothesis: *"Subjects mainly choose options that favor their favourite team and not options that are negative towards the performance of this favourite team".* The binomial tests gave a first insight into the existence of a potential *Home bias*. However, the paired sample t-tests shed light on the fact that this potential *Home bias* leads to bad predictions for some of the groups of subjects.

When comparing the performance of the four different algorithms on the percentage of correctly predicted outcomes, the confidence-weighted prediction method showed to be most effective. This algorithm was followed by the most-confident prediction method. However, for the reason that these methods could have given correct answers by chance, they were assessed using the following categorical correlation methods: The F1-score, Cohen's Kappa, and

Matthews correlation coefficient. All these measures showed that the *Surprisingly Popular* method underperformed compared to the most-confident prediction and the confidence-weighted prediction. However, the *Surprisingly Popular* method seems to give better predictions than the majority prediction method. In addition to these tests, the performance of the methods is compared using a McNemar Test.

As mentioned before, the subject group was divided into four groups for the McNemar test. These four different groups have been tested for similarity using a Kruskal-Wallis test. The results of the Kruskal-Wallis test gives no reason to assume that the four groups significantly differ on the variables age, gender, level of education, nationality, and subjective knowledge on the *Eredivisie*. The McNemar tests showed that there is a significant difference in performance for the *Surprisingly Popular* method when compared to the confidence-weighted prediction method. No significant difference in predictions was found when comparing the *Surprisingly Popular* method to the majority vote. The confidence-weighted prediction method outperforms both the *Surprisingly Popular* method and the majority vote.

Given these results using the variety of different measures, it is concluded that the *Surprisingly Popular* method does not perform better than confidence-weighted predictions. However, the results seem to show that the *Surprisingly Popular* method could be better performing than the majority vote. For this reason the second hypothesis: *"The Surprisingly Popular method gives more accurate predictions than averaging estimates of all subjects",* will not be rejected. However, the third hypothesis: *"The Surprisingly Popular method gives more accurate predictions than confidence-weighted aggregation methods",* will be rejected.

Based on the findings of this paper, it is concluded that the *Surprisingly Popular* method outperforms only the majority prediction method, and not the confidence-weighted prediction when the predictions are based on historical data of sports events. An explanation for the fact that the *Surprisingly Popular* method was not the most efficient algorithm in predicting the correct outcome by aggregating the predictions of subjects in a crowd, could be that this algorithm is actually not the most efficient, or that the methods used in this paper were not designed well enough to prove that the model is better.

Implications

As is shown in this paper, the *Surprisingly Popular* method give better predictions than the majority vote when groups of subjects judging whether statements based on historical football data are true or not. The *Surprisingly Popular* method did not outperform all of the

other methods as was expected. Seeing that these statements are based on a specific sport in a specific country, the generalizability of these results can be doubted. However, this paper has shown that the confidence-weighted prediction method had promising results and could be of interest in future research. The results of this paper could contribute to research into the improvement of aggregation models to improve *the Wisdom of Crowds*. Furthermore, some of the limitations and recommendations of this paper, which will be mentioned in the following sections, could give useful insights for follow-up research.

Limitations

In the progress of writing this thesis, it became clear at an early stage that there would be limitations in the experimental design of the research. It became apparent that due to Covid-19, it would be impossible to use actual bets on sports events. Therefore, it was chosen to continue the paper using the judgements of people on whether statements based on historical data of sports events were true or not. Likely, using historical data was not a perfect substitution for measuring correct predicting behaviour. Unfortunately, it was not possible to use predictions on actual events.

Furthermore, subjects in this study were not rewarded for choosing the correct answers, which made it less important to give correct answers. When/ there are no incentives for people to correctly answer the questions, it could be that they randomly picked answers and that they did not give well thought out answers. In addition to this, the phrasing of sentences could be of influence on the answers given by the subjects. Certain questions could have been phrased in a way that nudges people towards a particular answer. Some of the percentages asked could have been misleading, and it could possibly be that subjects do not think about the possibility of a draw between clubs in these matches. Also, the number of statements that were actually false could influence the performance of subjects. Seeing that eight out of the ten statements were actually false, this could make subjects more prone to question their own answers.

Another limitation of this study is the small number of statements which the subject had to rate as true or false. For this reason, some of the algorithms did not have give a value for all of the groups. As mentioned before, the most confident method was not applicable for all of the values. Also, the difficulty of the ten statements has not been assessed. It could be that some of the statements have been more difficult than others, or maybe even to difficult,

influencing the scores of the subjects. Lastly, in addition to the small number of statements, the number of subjects who actually completed the survey could have been larger.

Future recommendations

It is recommended that future research looks into the use of the *Surprisingly Popular* method in predicting actual outcomes of sports events instead of using historical data. The focus of this paper was on the Dutch football league, future research could look into other football leagues or even into other sports. Furthermore, it would be interesting if a new metacognitive question would be used, instead of asking people what percent of other subjects would choose a particular answer. This new metacognitive question could potentially contribute to a more precise and robust aggregation method than the *Surprisingly Popular* method. As the results of this research indicate that the confidence-weighted prediction method outperformed the other methods, this model could be a decent focal point in future works.

As mentioned in the limitations section, it would be advised to use larger sample sizes, and to question the subjects on more statements than the ten used in this paper. Also, it would be preferred if subjects are somehow incentivized to make correct predictions. This should be done to make sure that people put more effort in answering the questions in the survey. Furthermore, it might be interesting to look into other measures of a possible *Home bias* in betting, and testing whether the *Surprisingly Popular* method can help in situations when this bias is present. Lastly, seeing that the confidence-weighted method gave more accurate predictions than the *Surprisingly Popular* method, it could be interesting to use a confidence related question when coming up with a new metacognitive follow up question.

Bibliography

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370-374.

Brown, A., & Reade, J. J. (2019). The wisdom of amateur crowds: Evidence from an online community of sports tipsters. *European Journal of Operational Research*, *272*(3), 1073-1081.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267-280.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Franch, F. (2013). (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics*, *10*(1), 57-71.

Galton, F. (1907). *Vox Populi . Nature, 75(1949), 450–451.* doi:10.1038/075450a0

Görzen, T., & Laux, F. (2019). *Extracting the Wisdom from the Crowd: A Comparison of Approaches to Aggregating Collective Intelligence* (No. 56). Paderborn University, Faculty of Business Administration and Economics.

Herzog, S. M., & Hertwig, R. (2011). The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making*, *6*(1), 58-72.

Hulsen, T. (z.d.). *EredivisieStats - Alle uitslagen, standen en clubgegevens van de Nederlandse Eredivisie Voetbal*. Retrieved from http://www.eredivisiestats.nl/wedstrijden.php

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

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., ... & Couzin, I. D. (2018). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface*, *15*(141), 20180130.

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*(4), 322.

Lee, M. D., Vi, J., & Danileiko, I. Testing the Ability of the Surprisingly Popular Algorithm to Predict the 2017 NBA Playoffs.

Lock, S. (2019, September 23). Market value of online gambling worldwide 2017 and 2024. Retrieved from https://www.statista.com/statistics/270728/market-volume-of-online-gaming-worldwide/

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, *108*(22), 9020-9025.

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

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, *11*(3), 279-300.

McCoy, J., & Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. arXiv preprint arXiv:1703.04778.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, *13*(3), 279-301.

Staněk, R. (2017). Home bias in sport betting: Evidence from Czech betting market. *Judgment and Decision Making*, *12*(2), 168.

Surowiecki, J. (2004). The wisdom of crowds. 2004. *New York: Anchor*.

Trajman, A., & Luiz, R. R. (2008). McNemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scandinavian journal of clinical and laboratory investigation*, *68*(1), 77-80.

Appendix A tables and figures

Level of education:

Table 8

Highest level of education completed

	Frequency	%	Cummulative %
High School Diploma	45	28.5	28.5
Bachelor's Degree	87	55.1	83.5
Master's Degree	22	13.9	97.5
PhD	1	.6	98.1
Other	3	1.9	100
Total	158	100	

Binomial tests:

Table 11

Binomial tests four supporters groups

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (two-tailed)
PSV	Group 1	yes	16	.29	.50	.002
	Group 2	no	40	.71		
	Total		56	1.00		
Feyenoord	Group 1	yes	167	.63	.50	.000
	Group 2	no	97	.37		
	Total		264	1.00		
Ajax	Group 1	yes	137	.71	.50	.000
	Group 2	no	55	.29		
	Total		192	1.00		
FC Utrecht	Group 1	yes	25	.57	.50	.451
	Group 2	no	19	.43		
	Total		44	1.00		

Cohen´s kappa:

Table 12

Cohen's kappa majority prediction

		Actual		
		1.00	2.00	Total
Majority	1.00	2	7	9
	2.00	0	1	1
Total		2	8	10

Symmetric measures

		Value	Asymptomatic Standard Error	Approximate	Approximate Significance
Measure of Agreement	Kappa	.054	.064	527	0.598
N of Valid Cases		10			

Note. Using the asymptotic standard error assuming the null hypothesis.

Table 13

Cohen's kappa Surprisingly Popular method

		Actual		
		1.00	2.00	Total
Surprising	1.00	2	4	6
	2.00	0	4	4
Total		2	8	10

Symmetric measures

		Value	Asymptomatic Standard Error	Approximate	Approximate Significance
Measure of Agreement	Kappa	.286	.194	1.291	.197
N of Valid Cases		10			

Note. Using the asymptotic standard error assuming the null hypothesis.

Table 14

Cohen's kappa most-confident prediction

		Actual		
		1.00	2.00	Total
MostC	1.00	2	3	5
	2.00	0	5	5
Total		2	8	10

Symmetric measures

		Value	Asymptomatic Standard Error	Approximate	Approximate Significance
Measure of Agreement	Kappa	.400	0.232	1.581	.114
N of Valid Cases		10			

Note. Using the asymptotic standard error assuming the null hypothesis.

Table 15

Cohen's kappa confidence-weighted prediction

		Actual		
		1.00	2.00	Total
ConfidenceW	1.00	2	1	3
	2.00	0	7	7
Total		2	8	10

Symmetric measures

		Value	Asymptomatic Standard Error	Approximate	Approximate Significance
Measure of Agreement	Kappa	.737	0.241	2.415	.016
N of Valid Cases		10			

Note. Using the asymptotic standard error assuming the null hypothesis.

Paired-sample t-tests:

Table 16

Paired Sample t-test Ajax

	Paired Differences							
				95% CI of the Difference				
	Mean	Std. Deviation	Std. Error	Lower	Upper	t	df	Sig. (two-tailed)
Favourite-Other	-0.207	0.291	0.042	-0.291	-0.122	-4.92	47	.000

Table 17

Paired Sample t-test Feyenoord

	Paired Differences							
				95% CI of the Difference				
	Mean	Std. Deviation	Std. Error	Lower	Upper	t	df	Sig. (two-tailed)
Favourite-Other	0.198	0.334	0.041	0.116	0.280	4.82	65	.000

Table 18

Paired Sample t-test FC Utrecht

	Paired Differences							
				95% CI of the Difference				
	Mean	Std. Deviation	Std. Error	Lower	Upper	t	df	Sig. (two-tailed)
Favourite-Other	-0.159	0.267	0.081	-0.339	0.021	-1.97	10	.077

Table 19

Paired Sample t-test PSV

	Paired Differences							
				95% CI of the Difference				
	Mean	Std. Deviation	Std. Error	Lower	Upper	t	df	Sig. (two-tailed)
Favourite-Other	0.131	0.382	0.102	-0.089	0.351	1.28	13	.222

McNemar tests:

Table 20

Majority and Surprise

	Surprise	
Majority	True	False
True	14	12
False	4	10

Test Statistics of the McNemar Test

	Majority and Confidence
N	40
Exact sign. (two-tailed)	0.077

Note. Binomial distribution was used.

Table 21

Majority and Surprise

	Surprise	
Majority	True	False
True	13	0
False	3	22

Test Statistics of the McNemar Test

	Majority and Confidence
N	38
Exact sign. (two-tailed)	.250

Note. Binomial distribution was used.

Table 22

Majority and Confidence

	Confidence	
Majority	True	False
True	11	2
False	13	12

Test Statistics of the McNemar Test

	Majority and Confidence
N	38
Exact sign. (two-tailed)	.007

Note. Binomial distribution was used.

Appendix B Survey

Introduction page 1

This survey is part of a study for my Master thesis Behavioural Economics.
In this survey you are asked to rate whether certain statements are true or false.
The survey is completely anonymous and the results of the study are not made public. The information obtained from this investigation will be treated confidentially.

Answering the questions should only take 5 minutes of your time.

Thank you very much for filling in this survey!

Introduction page 2

You will be confronted with a total of 10 statements on Football statistics of clubs in the Dutch Football (soccer) league the *Eredivisie*.

The statistics are based on matches played in the last ten seasons of the *Eredivisie* (2009/2010 - 2019/2020, including al matches played up until today).

After each statement you are asked to rate whether the statement is true or not.
Following this question, you will be asked to give the level of confidence in your answer on whether the statement is true or false.
Lastly, you are asked to estimate the percentage of other participants in the study that rate the statement as true.

You do not need have to any knowledge of football or thi particular league to participate.

I ask you kindly not to look these statistics up on the internet.

Good luck!

Statement 1

Ajax (Amsterdam) has won more than 60% of its matches against AZ (Alkmaar) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

50

Statement 2

Feyenoord (Rotterdam) has won more than 30% of its matches against Ajax (Amsterdam) in the last 10 seasons,
(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 3

Ajax (Amsterdam) has won more than 50% of its matches against PSV (Eindhoven) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ○ True
- ○ False

Rate your level of confidence for your answer on the previous question:

- ○ Totally uncertain (50%)
- ○ Low confidence (60%)
- ○ Moderate confidence (70%)
- ○ High confidence (80%)
- ○ Very high confidence (90%)
- ○ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 4

PSV (Eindhoven) has won more than 65% of its matches against AZ (Alkmaar) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 5

Feyenoord (Rotterdam) has won more than 50% of its matches against AZ (Alkmaar) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 6

Ajax (Amsterdam) has won more than 50% of its matches against Fc Utrecht (Utrecht) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 7

AZ (Alkmaar) has won more than 50% of its matches against Fc Utrecht (Utrecht) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 8

Fc Utrecht (Utrecht) has won more than 30% of its matches against PSV (Eindhoven) in the last 10 seasons,

(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 9

Fc Utrecht (Utrecht) has won more than 30% of its matches against Feyenoord (Rotterdam) in the last 10 seasons,
(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Statement 10

PSV (Eindhoven) has won more than 45% of its matches against Feyenoord (Rotterdam) in the last 10 seasons,
(2009/2010 - 2019/2020, including all matches played up until today).

- ◯ True
- ◯ False

Rate your level of confidence for your answer on the previous question:

- ◯ Totally uncertain (50%)
- ◯ Low confidence (60%)
- ◯ Moderate confidence (70%)
- ◯ High confidence (80%)
- ◯ Very high confidence (90%)
- ◯ Certain (100%)

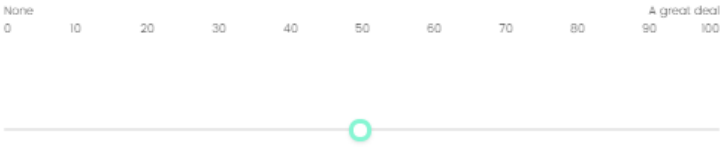
What percentage of other participants in the study do you think answered that the statement is true?

0 10 20 30 40 50 60 70 80 90 100

Percentage of other participants in the study

Control questions

How much knowledge do you have about Football?



Do you actively follow Football?

- ○ Yes, i often watch Football matches (at least once every two weeks pre-Corona)
- ○ Yes, i sometimes watch football matches (less than once every two weeks pre-Corona)
- ○ No, i never watch football matches.

Do you play Football or have you played Football in the past?

- ○ Yes, i currently play Football in a club context
- ○ Yes, i have played Football in the past in a club context.
- ○ No i have never played Football in a club context.

How much knowledge do you have on the Eredivisie?

Do you actively follow football games in the Eredivisie?

- ○ Yes, i watch Eredivisie matches often.
- ○ Yes, i sometimes watch Eredivisie matches.
- ○ No, i never watch Eredivisie matches.

Demographic questions

What is your age?

- ○ <18
- ○ 18-21
- ○ 22-25
- ○ >25

What is the highest level of education you have completed?

- ○ High school Diploma
- ○ Bachelor´s degree
- ○ Master´s degree
- ○ PHD
- ○ Other, namely:

What is your gender?

- ○ Female
- ○ Male
- ○ Other

Demographic questions

What is your nationality?

- ○ Dutch
- ○ Non Dutch, European
- ○ Other

Favorite Dutch football team in the Eredivisie

○ ADO Den Haag	○ PEC Zwolle
○ Ajax Amsterdam	○ PSV Eindhoven
○ AZ Alkmaar	○ RKC Waalwijk
○ FC Emmen	○ Sparta Rotterdam
○ Feyenoord Rotterdam	○ FC Twente
○ Fortuna Sittard	○ FC Utrecht
○ FC Groningen	○ Vitesse Arnhem
○ sc Heerenveen	○ VVV-Venlo
○ Heracles Almelo	○ Willem 2 Tilburg

Home city questions

In which city do you currently live?

- ○ City:

Which city do you consider as your home city?

- ○ City:

Ending page survey

This is the end of the survey

Thank you very much for filling in this survey!

If you have any questions or remarks you can contact me at:
435494ck@eur.nl