ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics
Master Thesis Behavioural Economics

Combining and de-biasing expert judgement

Name student: Rob van Schaik
Student ID number: 457439

Supervisor: Dr. T. Wang
Second assessor: A.C. Peker

Date final version: 03/02/2021

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

The wisdom of crowds is an effective way of reducing individual biases to get accurate predictions. When expertise is necessary to give an estimation for a certain problem no large samples can be obtained for a wisdom of crowds approach. These small expert samples can be combined with more advanced models that weigh individual predictions. Models that effectively achieve this are the Classical Model (Cooke & Goossens, 2008) and the Contribution Weighted Model (Budescu & Chen, 2014). However, their usefulness is limited by the types of data they require to function. Therefore, a new model is introduced that calculates expert weights to improve aggregate predictions that can be used with all types of numerical data.

This model calculates the relative accuracy of an expert's past prediction to the expert’s peers and weighs expert’s scores accordingly. Values of future uncertainties can then be calculated using a weighted combination of expert predictions.

The efficacy of the model was analysed using a total of 514 questions in 45 expert samples**.** Which exact specifications of the model optimize its effectiveness are calculated. On average, the model outperforms the best expert in the group and average predictions by 0.15 and 0.05 standard deviations, respectively. It is also shown that calculated expert weights are positively correlated with better predictions.

Future research is needed to determine how the model performs in practice and how the model’s usability compares to its alternatives.

Key words: wisdom of crowds; expert bias; normalisation; relative accuracy; weighted estimation

Table of Contents

1. Introduction

The mean guess of a large crowd is surprisingly accurate at making predictions and usually exceeds the accuracy of the smartest individuals in that crowd. First shown by Galton (1907), a large crowd can extremely accurately estimate the weight in an ox weighing contest. Likewise, combining the knowledge of more players causes a higher performance when playing the game show The Price Is Right (Lee, Zhang, & She, 2011). It can also be used to explain aspects of the stock market and betting markets and to depend policy on (Surowiecki, 2005). The wisdom of crowds, the combining of large samples, can be utilized to improve estimates regarding uncertain events. Nevertheless, several misconceptions exist concerning the wisdom of the crowds and expert judgements; people are often unable to recognize the increased accuracy of averaging values of different experts compared to their individual estimates (Larrick & Soll, 2006). This highlights the importance of the potential value there is in combining knowledge that is currently unrealised. This method can be applied in many fields of study, not only the financial or economic sector.

Combining values in forecasting has been applied in multiple disciplines and can potentially improve estimations (Clemen, 1989). The wisdom of crowds approach works well with large, unbiased, and random samples. In certain cases, random sampling will not be a logical approach to calculate values using a wisdom of crowds approach. Some estimations can only be done by subject matter experts; people without sufficient knowledge will then not be able to contribute valuable data. Bates and Granger (1969) show how two combined forecasts can outperform the better forecast of the two. This analysis does require the forecasts to be unbiased. The idea of giving different weights to forecasts is also introduced, instead of using a simple average. Namely, by increasing the weights for predictions with a relatively lower mean square error.

The wisdom of crowds is likely to increase with the size of the crowd. However, in several cases it is unlikely to find a sufficiently large sample to depend predictions on. Problems that require expertise from the respondents are defined by small sample sizes. It would therefore be beneficial to be able to obtain wisdom from small expert groups which is not affected by small sample sizes and biases. Still, from small samples wisdom can be extracted that exceeds the accuracy of the best expert in the group. Cooke and Goossens (2008) study the effectiveness of the wisdom of the crowds for cases that require expert judgement, such as volcanic activity and nuclear applications. By developing a model, they can improve the accuracy of combined estimates, even in small samples. It is important to improve the accuracy of the prediction of uncertain events in studies that require expert judgements due to the potential societal effects accurate predictions can have.

In certain fields experts have been criticized on their ability to make predictions in their field of expertise. Armstrong (1980) addresses this; even though experts have been unable to predict many important events in the past, people still seek their expertise. The Classical Model (CM) by Cooke and Goossens (2008) and the Contribution Weighted Model (CWM) by (Budescu & Chen, 2014) are wisdom of crowds models for expert data. These models combine expert judgements to create more accurate forecasts that outperform metrics such as the best expert and simple average prediction. Weighted predictions are used to combine expert judgement, similar to Bates and Granger (1969). In these models, weights are based on past performance. This way, the past is actively taken into consideration when estimating future events.

However, the CM relies on predicted confidence intervals and the CWM on specific probability data. It is at most uncertain whether experts are good at estimating confidence intervals accurately. Behavioural biases can affect estimates and confidence intervals (Soll & Klayman 2004). Therefore, a model that does not rely on confidence intervals could be less prone to behavioural biases. Additionally, solely using a single (average) estimation per expert would make a model easier to apply in practice.

This paper therefore aims to find a method to combine simple mean expert judgements by developing a new approach in combining expert judgements that increases the accuracy of the aggregate prediction. As well as to test a new model that only uses simple mean predictions and can be used on all types of numerical data, and to quantify to what extent predictions can be improved using a more practical weighted model. Thus, creating a model that can be used with all types of numerical data and does not require confidence intervals.

The paper is structured as follows. First, the current literature is reviewed. In the Methodology section the model used to combine expert judgements will be introduced and the analysis and the necessary data for the analysis will be described. Then, the results will be presented and analysed. Finally, the results will be evaluated in the discussion, followed by a general conclusion.

2. Literature review

Wisdom of crowds

Similarly to the aforementioned ox weighing contest, wisdom of crowds can be applied to more complex cases. In such cases, having a crowd with more subject knowledge does not necessarily improve the estimations. More complex cases, such as stock market predictions, intuitively require more expertise than simple wisdom of crowds examples where the amount of beans in a jar has to be guessed or an ox has to be weighed. Nevertheless, the expert crowds result in higher shared statistical errors, likely because experts have similar knowledge and judgement (Treynor, 1987). This is in line with the result that crowd performance decreases when individuals in a crowd have access to the predictions of others (Lorenz et al., 2011). Chen et al. (2014) investigates the wisdom of crowds in the stock market using opinions from people on the website Seeking Alpha. By investigating keywords in people's online posts their opinion is found to be a significant predictor of unexpected aspects of earning report releases. Correctly estimating these values could arguably be 'priced in', and not be a surprise to the market itself. However, significantly higher abnormal returns are also positively correlated with positive terms used on social media. Therefore, an aggregate of the crowds' opinion can make more successful stock predictions. This seemingly indicates that the crowd is wiser than the experts that make initial earning report predictions. Therefore, the question is when to ask a random crowd and when to consult experts. Cooke and Goossens (2008) argue that experts should be consulted when enough uncertainty exists within a field of expertise. Then, experts will likely disagree, there will be no structural bias, and statistical error is likely to decrease.

This does not mean that the wisdom of crowds always leads to more accurate predictions. Smaller samples can be affected by outliers, and if the entire group is affected by biases the mean results will also be (Simmons et al., 2011). Expert judgements can thus become inaccurate. Denrell and Fang (2008) investigate the accuracy of experts who have made impressive predictions in the past. They show that, even though people that have made unlikely predictions are often considered to be good experts, this is not the case. Having made bold predictions that have come true even seem to indicate an overall inaccuracy in terms of predictions. Not focussing on an individual but on a group of experts, therefore, has the potential to get rid of biases caused by recognition after making unlikely and true predictions.

Expert Bias

The most prominent bias that causes inaccurate predictions by experts is overconfidence. Overconfidence can cause confidence interval estimates to become too narrow. When asked to estimate 80 or 90 percent confidence intervals, on average, overconfident answers are given. However, more accurate confidence intervals are estimated when also asked for a median value estimation (Soll and Klayman 2004). Thus, inaccuracy in estimated confidence intervals is also affected by the framing of the question. The appearance of overconfidence could alternatively be caused because experts are averse to showing their actual uncertainty. When experts are incentivised, overconfident predictions decrease (Cesarini, Sandewall, & Johannesson, 2006).

Past success causes more confidence in the future. When analysing stock market analysts' forecasting predictions, successful past estimations cause analysts to become overconfident in their abilities (Hilary & Menzly, 2006). The potential influence of self-serving bias, where individuals praise themselves for their success and blame other elements for failure, is also addressed, but no conclusive evidence is found in the analysis. On the contrary, research shows that unsuccessful past predictions can cause bolder predictions in the future. More risk will then be taken to correct these mistakes by analysts (Evgeniou et al., 2013). Therefore, both a successful and unsuccessful record can cause, respectively, overconfident and bold predictions.

There are several additional biases that could logically affect expert judgement. Present judgements could be affected by trying to correct past imprecisions, and present predictions could be affected by the anchoring of past predictions. Meaning that the value of a new prediction can be affected by the values of old predictions (Tversky & Kahneman, 1974). Furthermore, experts might overestimate their own abilities by overestimating or misjudging their past performances. Hindsight bias causes the idea that previous uncertainties were already known (Fischhoff & Beyth, 1975). Even sophisticated subjects are generally unable to use disconfirming evidence to check predictions (Wason, 1968; Wason, 1969). This indicates that expert predictions may suffer because experts are inclined to only evaluate positive hypotheses.

Nevertheless, there is also evidence that expertise reduces bias caused by overconfidence. The Dunning-Kruger effect states that a low amount of knowledge causes a large spike in overconfidence, whereas an increase in knowledge reduces the overconfidence and makes people more aware of their personal abilities (Kruger & Dunning, 1999). This is, however, a relative measure; it is a relationship between confidence and expertise. Therefore, it does not argue that experts are immune to overconfidence biases.

It is hard to determine which biases play a role in practice when it comes to expert judgements and probability estimates, or if this even is the case. For example, overconfidence appears when experts are asked to predict confidence intervals (Soll & Klayman 2004). Additionally, when asked to predict wide lower and upper bound values, the results can be affected by the perception of these percentages.

Weighed wisdom of crowds

In the literature, a distinction between mathematical and behavioural methods of combining expert judgement exists. The Classical Model by Cooke and Goossens (2008) could be considered a mathematical method of combining expert judgements. Combining judgements in groups of peers can alternatively be done with behavioural approaches. To exemplify, communication between peers can be used in forecasting methods (Rohrbaugh, 1981; Rohrbaugh, 1979; Flores & White, 1989). Flores and White (1989) conduct an experiment where subjects predict the Dow Jones Index both individually and in groups. The combined forecasts are, on average, more accurate than a combined (unweighted) prediction. This paper will focus on a mathematical approach, using a combination of independent observations to improve estimates and reduce bias.

Different models have been designed in order to maximize the effectiveness of the wisdom of crowds approach. Cooke and Goossens (2008) use their Classical Model to combine expert judgements. On average, the CM effectively outperforms the values estimated by the best expert and the group average. The model uses a weighed system that calculates a combined estimation based on the performance of experts. The performance is measured using seed questions (or calibration questions); questions asked to the experts concerning their field of expertise to measure their performance. Meaning that experts must estimate a 5%, 50% and 95% value for each question. The percentages define an experts' perceived chance that the true value will fall below the 5%, 50%, or 95% quantile.

Confidence estimates are used to calculate the accuracy of the predictions, and the size of the individual confidence intervals to calculate precision. The product of accuracy and precision determines the final score for each expert. The precision calculation is used to determine confidence for each expert. This measure's score increases with more narrow confidence intervals. However, if larger precision causes more incorrect answers the algorithm generally causes the overall score to decrease. Estimating narrow confidence intervals is a sign of confidence in an individual's personal ability to predict outcomes. The combination of the two elements therefore accounts for overconfidence. Additionally, the estimation technique that includes the estimation of a mean value, which improves the accuracy of estimated confidence intervals, has been shown to reduce biased estimation (Soll & Klayman, 2004).

The CM has several disadvantages. It does not distinguish between accurate and slightly accurate answers. It uses a method similar to a chi-squared test to calculate accuracy of predictions of all the answers to seed questions. A chi-squared distribution is used for four intervals: two for values above and below the 5% and 95% quantile, and two for values between 5% and 50% and the 50% and 95% quantiles. Therefore, it is only decisive for estimates to be within the confidence interval, it does not matter how close the estimate is to the true value. Additionally, slight over and underestimations can cause individual scores to become lower than they maybe should be. For example, a relatively good expert that slightly overestimates values could have most of his estimates in the upper bound of his confidence intervals. Therefore, a relatively good expert could have a large reduction in score caused by the CM due to over precise estimations. Moreover, the model often causes individual experts to get a weighted score equal to zero. This could possibly result in an increase in statistical error which is generally decreased by larger sample sizes.

There is some additional uncertainty regarding the CM and its effectiveness. Clemen (2008) elicited expert scores using the CM and out of sample data. Using these data resulted in different results than the original paper that used the TU Delft Structured Expert Judgement (SEJ) dataset. It even resulted in an adverse effect of the CM; it would only improve predictions in 40 percent of cases compared to the mean. In this analysis, a simple average outperforms the weighted score method. It is therefore questionable whether the method is effective in practice. The TU Delft SEJ data seems to have a wider variety than the data used by Clemen (2008), but the results, nevertheless, indicate uncertainty regarding reproducibility using different data. Certainty and consistency are important when eliciting expert wisdom since decisions that require certainty depend on the outcomes. Otherwise, policy makers could become reluctant to use (weighted) wisdom of crowds techniques in practice.

Another model is the Contribution Weighted Model, used for probabilistic judgement (Budescu & Chen, 2014). This model, similarly to the CM, calculates a weight for individuals of the crowd based on their performance. Budescu and Chen (2014) argue that the CM tends to overweight certain experts, which can cause the model to calculate extreme outcomes. Therefore, the authors argue that a model which calculates expert weights based on the performance of the group when the experts is included, compared to when the expert is not included, would be an effective method of calculating expert weights.

The CWM uses probability data. The model focuses on binary events and probability estimates whereas the CM is used for continuous variables with confidence intervals. Meaning that experts have estimated an expected likelihood for an event to occur. The quadratic scoring rule is used to measure the aggregated predictive accuracy of the experts. The following formula is used to calculate the performance of the crowd for event *j*:

$$(1)\ S_j = a + b \sum_{r=1}^{R_i} \left(o_{jr} - m_{jr}\right)^2$$

Where *a* and *b* are constants and o_{jr} and m_{jr} the binary indicator of the true outcome and the unweighted prediction of the crowd, respectively, for each event *j* (j = 1,..., J) and each outcome *r* (r= 1,..., R). The performance of the crowd is then calculated using:

$$(2)\ S = a + b \sum_{j=1}^{J} \left(\sum_{r=1}^{R_i} \left(o_{jr} - m_{jr}\right)^2\right)$$

Afterwards, for every single expert the overall score of the group is measured with and without their contribution to calculate their expert score (C_i). Thus, performance is based on the individuals of the crowd including and excluding individual experts. Only positive scores are used as expert weights to calculate final predictions with. Therefore, experts are rewarded a higher score for providing a positive contribution to the crowd. Formula (3) is used to calculate the input of each individual *i* (i = 1,..., N).

$$(3)\ C_i = \sum_{j=1}^{J} \frac{s_j - s_j^{-j}}{N_i}$$

There are some potential limitations to the CWM. Most notably, it can only be applied to a limited set of (probability) data. Additionally, it is unclear whether largely inaccurate predictions that converge a group average closer to the mean should cause a high individual expert score.

The efficacy of aforementioned models, the Classical Model and the Contribution Weighted Model, are analysed in their respective papers. Since the CWM (Budescu & Chen, 2014) is used for probabilities, a homogeneous scale from zero to one is used. Therefore, relative performance of the model can be expressed on this scale. The Classical Model (Cooke & Goossens, 2008) uses data on a variety of scales. Therefore, a different approach must be taken to test the model. This is done by using the model to calculate weighted predictions, then comparing the output to the individual experts. In 27 out of 45 cases the weighted prediction outperformed the unweighted prediction and the best expert. The model is also tested in practical situations, such as predicting the AEX and on real estate risk; where the true outcomes are within the confidence intervals predicted by the model.

Cooke and Goossens (2008) address in their conclusion how eliciting expert judgement using seed questions is both expensive and time-consuming. This means that a choosing for a better estimation becomes a trade-off between costs and accuracy. If expert opinions using only a single value can be used to calculate expert weights and improve a prediction it would be possible to use simple past predictions to apply to future predictions. This would make the model easily implementable and presumably less expensive to apply.

There are also more general criticisms on expert elicitation methods. Morgan (2014) argues that expert elicitation is not an effective method to apply in research. For example, when experts base their opinions on different models or methods, combining judgements could lead to inaccuracy. Alternatively, when making judgements, experts are unconsciously susceptible to biases. The availability heuristics and the anchoring and adjustment heuristic (Tversky and Kahneman, 1974; Kahneman, Slovic and Tversky, 1982) therefore affect predictions.

In summary

A wisdom of crowds approach can be a useful method to improve predictions. In certain cases, however, expertise is required since too much uncertainty exists for randomly chosen subjects to make predictions. But even experts suffer from biases when making predictions. Therefore, weighted models can be used to reduce biases and optimize combined predictions in small expert samples.

The CWM and CM effectively outperform metrics such as the best expert of the group and the simple average, as was analysed in their complementary papers (Budescu & Chen, 2014; Cooke & Goossens, 2008). This indicates that weighted models have potential to improve predictions. Therefore, the hypothesis is that a weighted model will increase performance even if only mean predictions are used. This paper does not necessarily aim to develop a method that will improve the accuracy of current models but to develop an effective model that does not require experts to define complex confidence intervals or give predictions that require binary outcomes. Thus, minimizing the possible effects of biases that come with the estimation of probabilities.

3. Methodology

3.1 Relative accuracy model

In order to combine expert judgements a model will be used that weights experts' scores performance relative to the group of experts. The intuition of the model is as follows. In general, experts get rewarded a higher score for outperforming their group of peers and a lower score for underperformance. This assessment is based on relative performance. If the entire group of experts performs poorly for a question, a worse performance will result in a relatively lower reduction of score. Similarly, when the entire group of experts is accurate, a more accurate result will only result in a slight increase in expert score. Moreover, when an expert performs contrary to the group average, the effect on the score will be larger. An accurate prediction will result in a higher score when the group average is inaccurate. Likewise, an inaccurate prediction causes a large decrease in score when the group average is accurate. The following equation is used to calculate expert scores (S_i) for the individual expert *i*:

$$(4)\ S_i = a + \frac{b}{Q} \sum_{q=1}^{Q} (|G_q - T_q| - |I_{iq} - T_q|)$$

Where Q (q = 1,...,Q) is the total amount of questions, and constants *b* and *a* respectively affect the multiplier of the calculated score per question and the initial score per expert. Since different estimates require different scales, no absolute numbers can be used to calculate expert scores. The relative difference from an individual expert estimate (I_{iq}) to the group average of experts (G_q), in terms of the true value of the estimate (T_q), is normalized to be comparable between different types of questions. This is done by calculating the difference in terms of a z-score; the statistical difference of a prediction from the mean of the group. The calculation of the z-scores for individual estimations per question (Iz_{iq}) and true values per question (Tz_q) are shown in formulas (5) and (6), respectively. Where (E_{iq}) is the estimation made by expert *i,* (R_q) the real value, and (M_q) the mean for question *q*.

(5) $Iz_{iq} = \frac{E_{iq} - M_q}{\sigma_q}$

(6) $Tz_q = \frac{R_q - M_q}{\sigma_q}$

The values Iz_{iq} and Tz_q are subsequently used to calculate a corresponding normalised value using a cumulative normal distribution using formulas (7) and (8). Where ϕ is the cumulative normal distribution function. This method has two advantages. Firstly, the scores are scaled from 0 to 1 using a cumulative normal distribution. To exemplify, average predictions are equal to 0.5 and values that largely exceed the mean prediction approach a value of 1. Secondly, the normalization takes the size of the standard deviation of a question into account. The data, in most cases, will not follow a normal distribution and therefore not satisfy the assumptions for the calculation of p-values. The scores themselves should therefore not be interpreted and should solely be used to enable the possibility of normalizing all expert judgements to a scale from 0 to 1. Since a scale from 0 to 1 is used for the calculations the group average (G_q) equals 0.5 on this scale. What the ideal values are for the constants *a* and *b* will be analysed (see analysis section).

(7) $I_{iq} = \phi(Iz_{iq})$

(8) $T_q = \phi(Tz_{iq})$

In formula (4), the absolute difference between the group average and the true value of a prediction ($|G_q\text{-}T_q|$) calculates the distance of the true value from the group average. The absolute difference between the individual estimate and the true value of a prediction ($|I_{iq}\text{-}T_q|$) calculates how much an individual experts' estimate differs from the true value. The difference between the two aforementioned aspects of the model therefore calculates how much closer an individual expert is to the true value than the group is. Since the values T_q and I_{iq} are measured on a normalized scale from 0 to 1, the minimal score for every question is -0.5 and the maximum score is 0.5. The constant *b* allows for weighing the scores for each question and can therefore be used to increase or decrease the magnitude for each question. Average scores per question of 0 (estimating the exact average of the group prediction) will result in an average weight of constant *a*. The score for every individual question is summated and divided by the total number of questions (Q). Negative scores are only obtained when experts repeatedly underperform. Negative expert scores are excluded from the final aggregation to prevent the algorithm from using estimations by underperforming experts as adverse values in calculating the final estimation. The score weights (S_i) for each individual *i* can then be used to combine expert weights with individual expert estimations (E_i) using the following formula:

$$(9)\ Weighted\ estimation = \sum_i^N (S_i * E_i) / \sum_i^N (S_i)$$

Similarly to the CWM (Budescu and Chen, 2014), this model uses a relative scoring measure, where predictions that add to the accuracy of the group are rewarded. However, there are some notable differences. This model solely needs mean estimates and can therefore be applied to all types of numerical data. Additionally, the model uses the relative performance of experts compared to the group. Therefore, it is measured how accurate an expert is relative to the group, instead of how much he improves the overall performance of the group. The CWM would reward experts with estimations that are equally wrong as the rest of the group and that would shift the mean estimate closer to the true value. The relative accuracy model only rewards experts if their predictions are relatively more accurate. This could improve predictions since accuracy and consistency determine the expert scores.

3.2 Normalised Contribution Weighted Model

The relative accuracy model has several differences compared to the Contribution Weighted Model (Budescu & Chen, 2014). By applying the new method of normalisation on the CWM a comparison between the two methods can be made. In the classical wisdom of crowds approach a large number of predictions (ideally by a randomly selected sample) are combined to get an accurate average value. The intuition behind the CWM is more in line with the original wisdom of crowds theory because the values of scores are determined by contribution to the group. For expert judgement with small sample sizes it is, however, unclear what the best approach is. The model formulated in equation (4) can be adjusted to let it depend on contribution instead of relative accuracy. In formula (10) the relative improvement of the group due to contribution by expert *i* is calculated. Where *a* and *b* are the previously used constants, (M_q) the mean and ($M_{q\text{-}i}$) the mean excluding expert *i* for question *q*. The exact same formulas cannot be used on these data since they do not include binary events. Therefore, a method similar to the score calculation in formula (4) is applied, using the intuition from the CWM. Subsequently, the final combined estimates are calculated using formula (9).

$$(10) \qquad S_i = a + \frac{b}{Q} \sum_{q=1}^{Q} (|M_q - T_q| - |M_{q-i} - T_q|)$$

It should, however, be noted that this formula is still different from the CWM. Since different types of data are used the models cannot be compared directly. This method allows for a general comparison between the two approaches; it is used to determine which approach results in better estimations. Which would give information about how the different methods in expert weight calculations compare. Therefore, it cannot be used to establish how the two models compare in overall accuracy.

3.3 Data

The TU Delft Structured Expert Judgement data is used to analyse the effectiveness of combining expert judgements. The same dataset was used by Cooke and Goossens (2008) to test the Classical Model. The data consist of 45 different expert judgement datasets from 2006 and later. This includes a wide variety of topics, such as healthcare, volcanic activity, and space debris. Seed questions are included with their corresponding true values. The amount of seed questions per topic varies between 7 and 18, the mode is 10 seed questions. The data also include real predictions from experts that have not been realised. These values will not be used in the following analysis since the results cannot be tested without the true values. Therefore, the seed questions can be analysed to see how accurate an expert was when he estimated the seed questions values. This allows for the possibility to evaluate the true accuracy of a large part of the predictions in the data.

Seed questions are measured either on a uniform or a logarithmic scale in the datasets. The logarithmic scales are used in certain cases to reduce the effects of outliers. Larger outliers increase the size of standard deviations which increases the variation in the calculation of expert scores. Therefore, the scales will not be changed for the use and analysis in this paper.

It is, however, still uncertain how true predictions compare to seed questions in terms of accuracy. Whereas true predictions are incentivised, they are also prone to biases. Such as the aforementioned bias where experts become overconfident after successful predictions (Hilary & Menzly, 2006), or where more risk is taken to compensate incorrect predictions (Evgeniou et al., 2013).

There are 45 different datasets with a wide variety of fields of expertise. Therefore, the data are diverse which should account for field specific biases. Additionally, there are a total of 514 questions with their true values to be analysed and a total of 463 experts. Given the large variety of the data and the quantity of the questions it is expected to be able to calculate relatively robust estimations.

3.4 Analysis

The accuracy of this model will be determined by calculating the accuracy of the weighted predictions in the TU Delft SEJ dataset. The results will be compared with the best expert of each group and the simple average (unweighted estimation) to determine if this method enhances the accuracy of the estimations. The best expert will be determined by calculating which expert has the highest expert score (S_i) using formula (4).

The model is tested for every question in a dataset with an estimated value and a corresponding realised value. The purpose of the model is to combine and estimate unrealised estimates. Therefore, the model can be tested using past predictions including realised values. The individual scores will be calculated per question in a dataset and will subsequently be combined using the aforementioned algorithm. For each question in an expert group that includes a realised value, the remaining questions will be used to calculate expert weights with. These weights will then be used to calculate an estimation for that question. To exemplify: for question 1 in a given dataset, all questions except for question 1 will be used to calculate expert scores. These scores will then be used to calculate what the aggregate prediction for question 1 would be. This value can be compared with the realised value of that question, the average estimation, and the best expert's estimation. This will be done for every question.

Considering that predictions for different questions are done on different scales, the estimations by the algorithm, the unweighted score, and the best expert, need to be normalised. This is done by calculating the relative amount of standard deviations difference from the true value. So, standard deviations are calculated separately per question, and are used to calculate how the three prediction methods perform in terms of difference from the realised value. Formula (11) will, for each question *q*, be used to calculate the relative difference (D_q) between a prediction (P_q) and a realised value (R_q), using standard deviations in an expert sample (σ_q). Small samples and diverging predictions can cause the standard deviations in a sample to become large. Large standard deviations would cause the results of the model to overestimate the accuracy of the model. Therefore, seed questions with extremely large standard deviations will be excluded from the performance tests. In the analysis the threshold is a standard deviation which is 50 times larger than the true value of the corresponding prediction. This is an arbitrarily chosen number that causes the extremely overestimated values to be excluded from the model while still including a large number of observations to calculate performance with.

$$(11) \qquad D_q = \frac{|R_q - P_q|}{\sigma_q}$$

Additionally, the formula that is used to calculate expert scores per question (formula (4)) contains two constants that allow for modification of the expert weight calculations. Different values of the two constants will be tested to optimize the effectiveness of the model. The optimal values for constants *a* and *b* will be determined by calculating the accuracy of the model using different values of these constants. For the score multiplier (constant *b*), values from 0.5 to 15 will be tested. For the initial value (constant *a*), values 0, 0.5, and 1 will be tested. The lower the value of *a*, the more experts with low scores will be excluded from the prediction. When *a* is equal to zero, only experts with a positive performance will be taken into account. Therefore, all experts with a relatively negative score will not be included in the final estimation of a question.

Furthermore, the performance of the normalised CWM and the relative performance model will be compared. Since the expert scores for a certain question are measured using seed questions it is also unclear if positive contributions (which can be relatively inaccurate predictions) also translate into better predictions in other questions. Therefore, it will also be analysed whether contribution or relative accuracy is more effective in determining expert scores. Finally, the effect of the number of seed questions and the number of experts on the accuracy of the predictions will be analysed. The data consists of 45 different datasets with alternating amounts of seed questions, ranging from 7 to 18. More seed questions or more experts means that more information is used to calculate expert weights. Therefore, it will be analysed to what extent these values improve the performance of the model.

4. Descriptive statistics

In the TU Delft SEJ dataset used in the analysis, the best expert outperforms the unweighted mean in 51.8 percent of the questions. This is not a statistically significant difference. Therefore, the mean and the best expert approach seem to perform reasonably equally.

In figure 1 can be seen that, for lower scores of *b*, the algorithm can significantly be closer to the actual value of a prediction than the simple mean on average. Furthermore, the difference decreases when the value of *b* is increased. This indicates that increasing the weight given to each question has a negative effect on the accuracy of the model.

Interestingly, the weighted average seems to outperform the unweighted average for a low value of *b*. This difference is decreasing with an increase in *b*. The opposite can however be observed when comparing the weighted score and the best expert. This can be seen in figure 2. Where, to a certain extent, an increase in *b* seems to improve the estimations of a weighted average relative to the best expert.

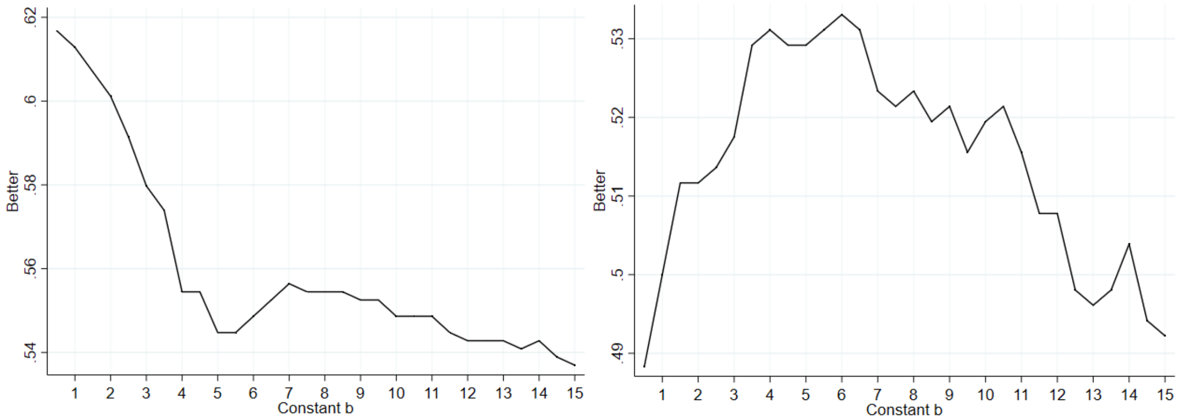


Figure 1 and 2. Equal weight compared to weighted predictions using different values for constant *b* and *a*=1. The graphs show the percentage of cases where the model outperforms the unweighted value (1) and the best expert (2).

A lower value for *b*, in general, means that the model will calculate less extreme expert weights, therefore it is possible that a low *b* causes small differences in estimations. Similarly, increasing *b*, thus increasing the differences in weight that will be calculated, seems to cause a relative improvement when comparing the model to the best expert approach. Since these measures are solely comparing which approach is closest to the true value of a question, it needs to be analysed how these different measures compare in terms of performance.

The variations in expert scores per dataset are visualised in figure 3. Since constant *a* is set to zero, this figure shows the calculated difference in scores for all experts. In the majority of cases, on average, experts have slightly negative scores. Negative scores will be excluded in the final aggregation of weighted predictions. Since higher values of constant *a* can be used in the calculation of expert weights not all low scores need to be excluded. A higher value for constant *b* will increase the number of experts to be excluded. A lower number of observations causes the wisdom of crowds effect to decrease. This could potentially explain the decrease in performance for higher values of *b* in figures 1 and 2. The effects of different constants and the performance of the model will be analysed in the results section. Furthermore, there are datasets with outliers. In these cases, there are experts that score significantly lower or higher than their peers.

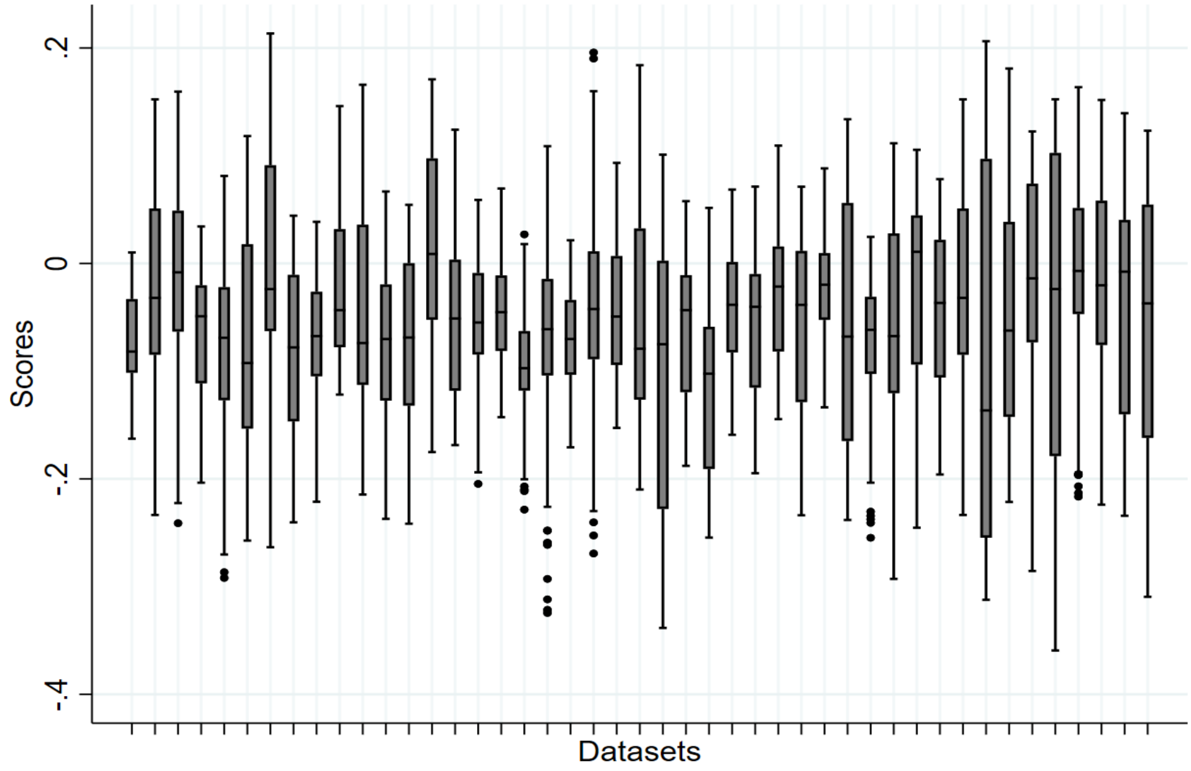


Figure 3. Expert scores for a=0 and b=1. Average scores and standard deviations are shown using boxplots for each of the 45 datasets. Dots depict outliers.

5. Results

Results are determined by calculating individual expert scores for each seed question using the remainder of the questions. So, in a dataset with ten seed questions, for all questions individual expert scores are calculated using the remaining nine questions. The results section is divided into separate parts. Firstly, the calculated expert scores and the differences between the 45 datasets will be analysed. Then, the impact of different values of the constant variables of the algorithm will be addressed. Afterwards, the general performance is evaluated. The differences in intuition of the current model will be compared with the intuition used in the

Contribution Weighted Model (CWM) (Budescu & Chen, 2014). Finally, the effects of more seed questions and experts in a dataset and the value of expert scores on the accuracy of the model will be examined.

5.1 Model performance

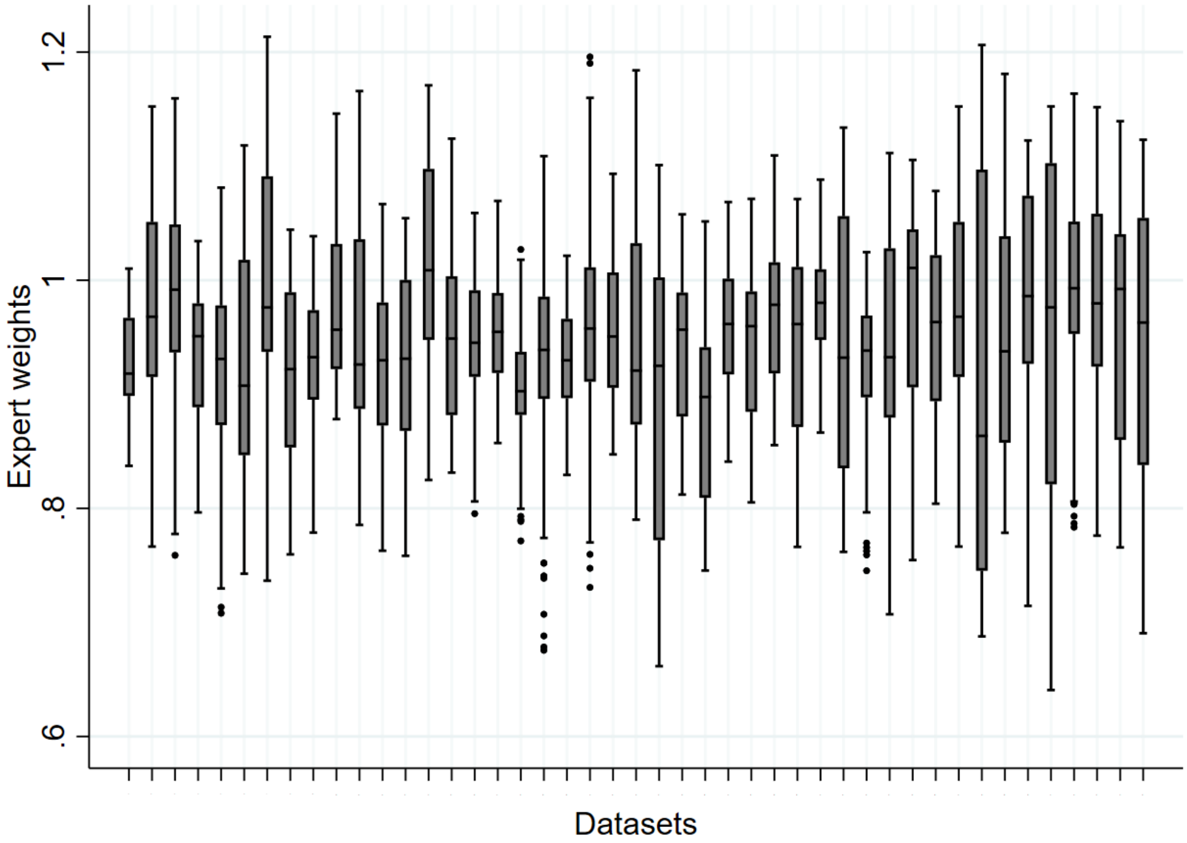


Figure 4. Expert weights for a=1 and b=1. Average scores and standard deviations are shown using boxplots for each of the 45 datasets. Dots depict outliers.

Expert scores

The expert weights which are used to determine final predictions with can be seen in figures 4 and 5. In these figures, the scores are visualised using separate boxplots for each of the 45 datasets. Expert weights have an average of 0.96 and have a minimum and maximum of 0.64 and 1.21 when constant *b* is equal to 1 (figure 4). Logically, the variation increases with an increase in the constant *b.* This constant magnifies the effect of divergence from the true value (relatively to the rest of the sample). Therefore, the variation increases with an increase in *b*. The expert weights in figure 5, where *b* is equal to 10, have an average value of 0.70, with a minimum of 0 and a maximum of 3.13. For this value, 24.5 percent of experts are excluded from the final estimation (meaning their weight equals zero). Similarly, a lower value of *a* would make the individual expert weights more likely to reduce to zero. This is discussed in the section below.

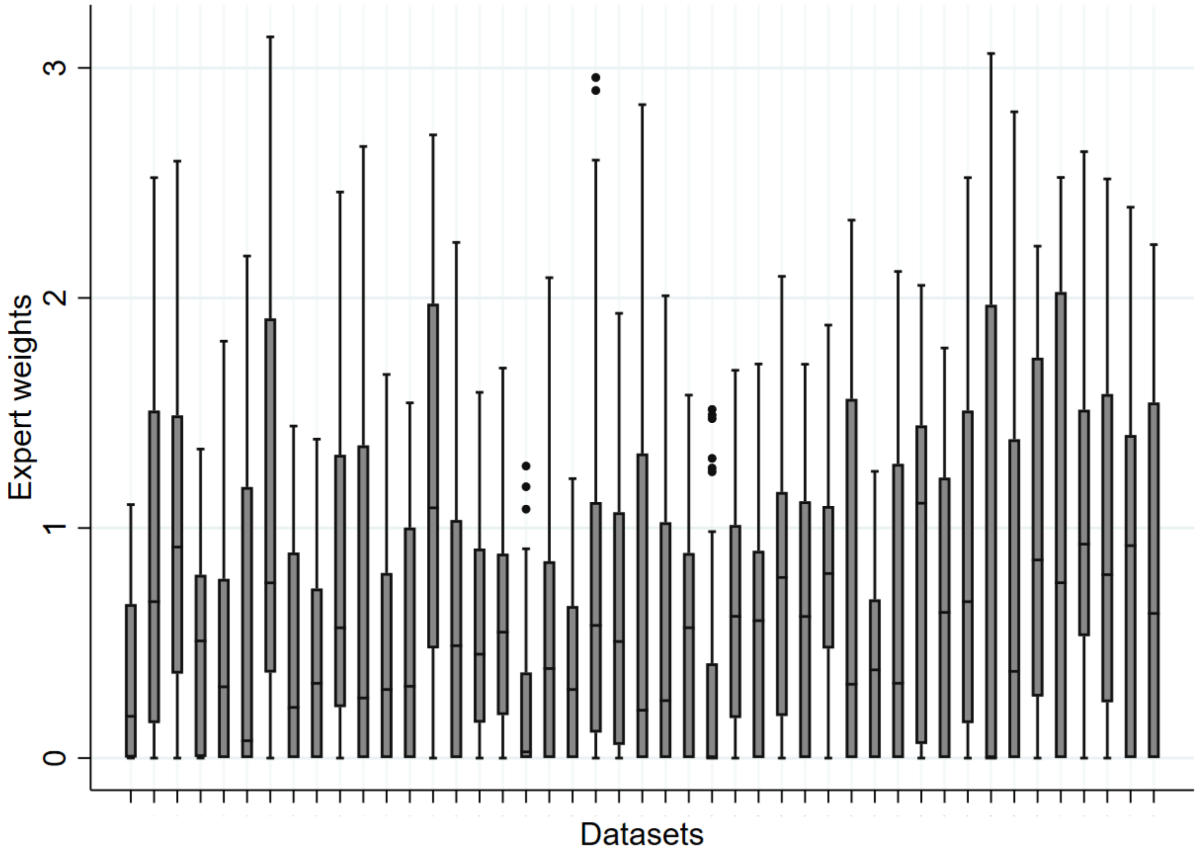


Figure 5. Expert weights for a=1 and b=10. Average scores and standard deviations are shown using boxplots for each of the 45 datasets. Dots depict outliers.

Examination of constants a and b

The formula used to calculate individual expert weights includes two constants. Formula (4) includes constants *a* and *b.* Constant *a* is used to set the range of the scores. If scores are below 0, they are excluded from the weighted estimation. Therefore, a lower value of *a* makes it more likely for underperforming experts to be completely excluded from the final estimation. A higher value of *a* will, therefore, make it more likely for experts to be included with a relatively lower weighted score. The effect that a change of *a* has on the accuracy of the model needs to be analysed.

The nature of the data makes it difficult to concretely state the effectiveness of the algorithm compared to the equally weighted predictions and the best expert approach. Therefore, the performance will be calculated in terms of standard deviations. These are determined separately for each question using the estimations of all experts that have answered a question.

Different values for constant *a* are examined by analysing the performance of the algorithm using different values of *b* used in formula (4). The results are visualised in figure 6, where the performance of the model is compared for different values of constant *a*. A lower value in this figure means that predictions are, on average, closer to the true value. A difference between the values a=0.5 and a=1 yields no significantly different results for all question weights (values of *b*). This indicates that if *a* is not low enough to reduce scores to zero, no effect can be observed. For a=0 it is the case that many expert weights become equal to zero. In this case, the effect is that the performance is lower for all values of *b*. Therefore, constant *a* will henceforth be equal to 1 to optimize the performance of the algorithm.

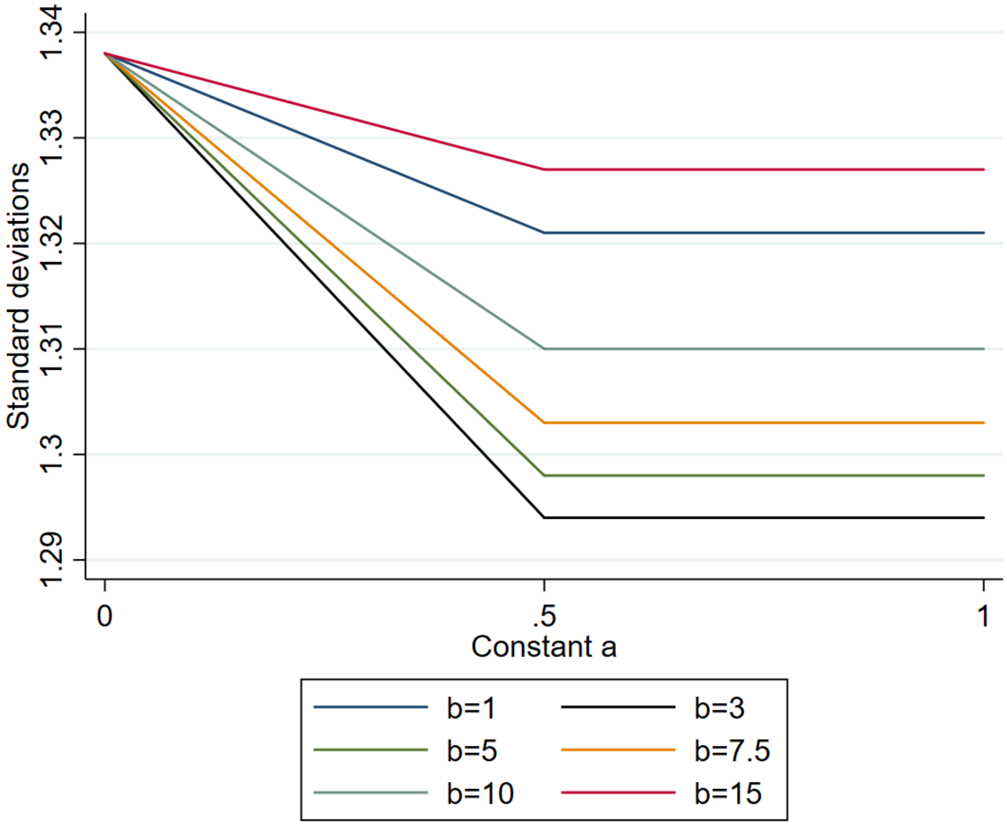


Figure 6. Difference in terms of standard deviations away from the true values for different values of constant *a*. The lines depict the difference in performance of the model due to a change of constant *a.* For *a* the values 0, 0.5, and 1 are used, as can be seen on the x-axis. The y-axis shows the difference between the estimated values by the algorithm and the realised values in standard deviations. Six values of *b* are used in the comparison.

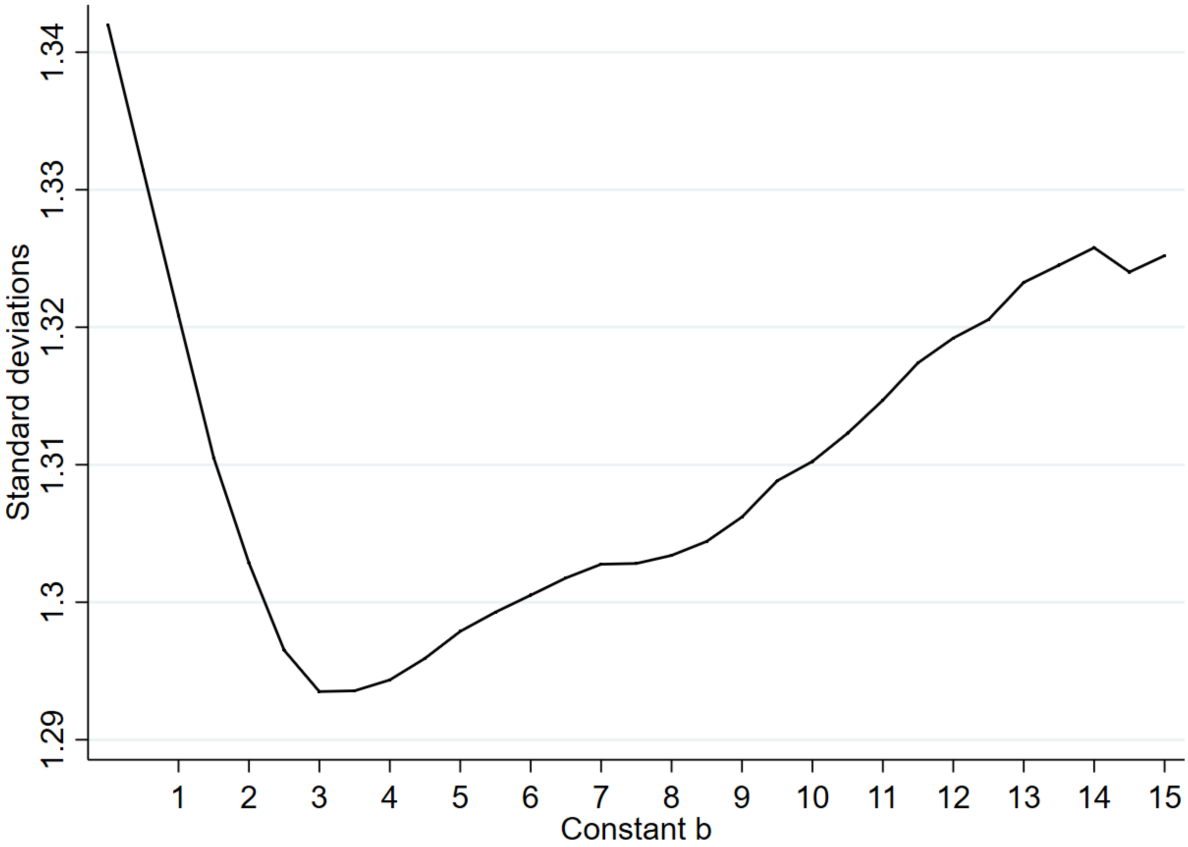


Figure 7. The average difference between the estimations by the model and true values in terms of the standard deviations difference (a value closer to zero means the model gives an estimation closer to the true value). This is calculated for all values of constant *b* between 0.5 and 15.

Scores for all values of *b* between 0.5 and 15 are calculated and shown in figure 7. The optimal value for *b* can be determined by choosing the value for which the model estimates scores that are relatively closer to the true outcomes. Which is the case when constant *b* equals a value between 3 and 4. For the value of 3 of the constant *b,* on average, the model predicts and estimation 1.29 standard deviations from the true value. To further illustrate the performance for different values of constant *b* the estimations by the model are compared to the mean and the best expert in figure 8. An optimal relative performance is again achieved when constant *b* approximately equals 3. In this case, the model outperforms the calculated mean estimation with 0.05 standard deviations, on average. Additionally, it outperforms the best expert by 0.15 standard deviations, on average. Both values are significant at a 1 percent level.

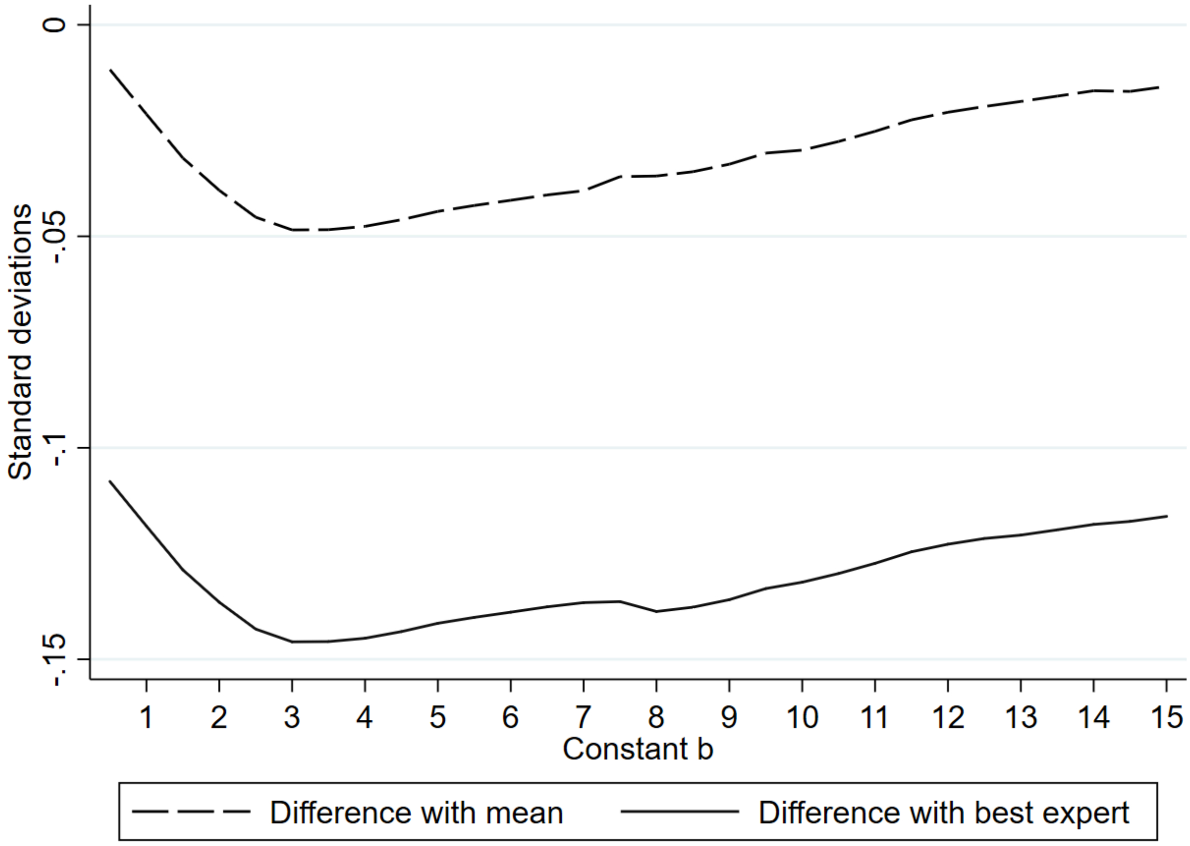


Figure 8. The average amount of standard deviations the estimation by the model is closer to the true value than the mean and best expert (lower is better). This is calculated for all values of constant *b* between 0.5 and 15.

Performance

The following three measures are used to determine the performance of the model: the estimations by the relative accuracy model, the unweighted average, and the best expert. The weighted aggregation is calculated by the new model, the unweighted average is the mean estimate of all experts per question, and the best expert is the prediction of the highest performing expert. To examine whether the new model improves predictions it is compared to the last two benchmarks.

On average, using the optimal value of constant *a* (*a*=1) and *b* (*b*=3) this model will estimate values that are 5 percent of a standard deviation closer to the true value than using an unweighted model. Similarly, estimations are 15 percent of a standard deviation closer to the truth than using the best expert in a group (see figure 8). Both are significant at p<0.01. A simple comparison between the performance of the model and the weighted score shows an outperformance of the model in 57.5 percent (284/494) of questions after deleting large outliers. Compared to the best expert, the same is the case in 53.8 percent of questions. These percentages simply indicate in how many cases the model's predictions are closer to the true value. They should therefore not be used to interpret the relative performances of the different approaches.

The relative performance of the model is compared to the mean and the best expert. This is visualised in figures 9 and 10, using the measure of standard deviations difference from the true value. The number of standard deviations the three measures are away from the true values of their corresponding estimates are used to visualise how well the predictions by the model perform in comparison to the mean and best expert approaches. The average distance to the true value in standard deviations is shown for 41 datasets, four outliers were excluded to improve the visibility of the results. The results show that the model is in most cases closer to the true value than both the mean and best expert. Therefore, this indicates that the relative performance tends to outperform the unweighted average and the best expert. Nevertheless, there are still many cases where the mean and best expert outperform the model. On average, the model outperforms the mean and the best expert in 60 percent of datasets (including the four outliers). Additionally, as was also shown in figure 8, the difference between the model and the mean is on average small (0.05 standard deviations).

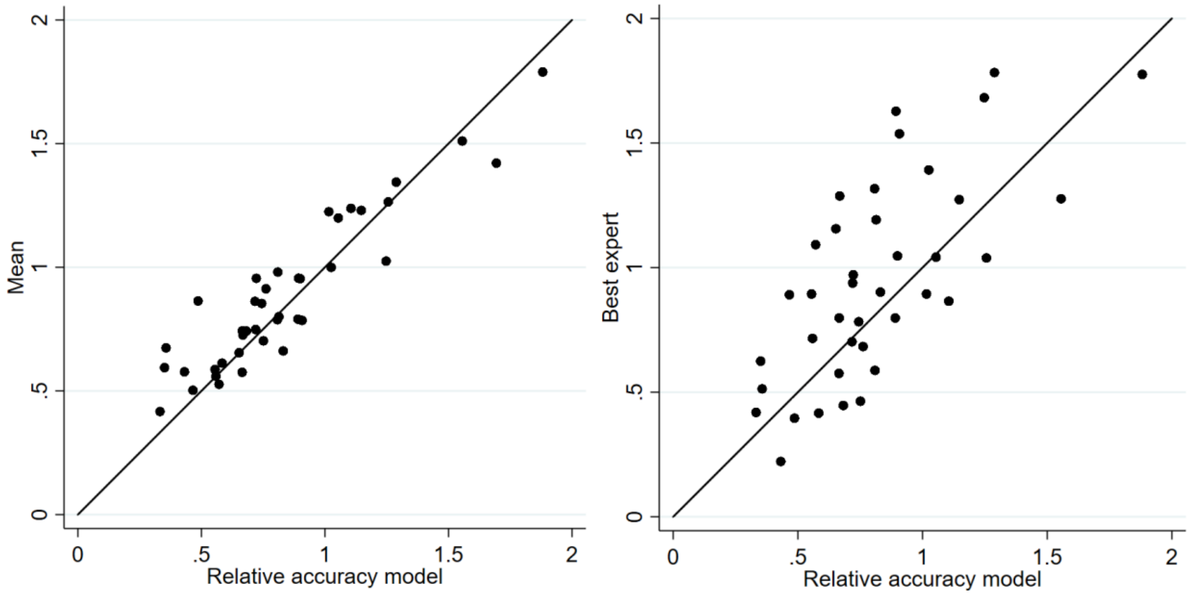


Figure 9 and 10. A comparison between the accuracy of the relative performance model (on the x-axis) and the mean estimations (figure 9) and the best expert (figure 10) (on the y-axis). Accuracy is measured by the number of standard deviations the estimations are from the true value (a value closer to zero means the model gives an estimation closer to the true value). Points above the line depict a higher performance by the relative performance model. The values are calculated per dataset and outliers are excluded. Constants: a=1 and b=3.

The standard deviations are relatively high. To illustrate this: in 43 percent of questions the standard deviation is larger than the true value. This is caused by the relatively low number of observations in the Structured Expert Judgement datasets. This number is also slightly increased by realised values that are close to zero.

Due to large standard deviations the final performance is affected by outliers. In the aforementioned results the overly large outliers are excluded. Where the standard deviation is more than 50 times larger than the true value (which excludes 34 out of 514 observations). This is an artificially chosen number that causes the exclusion of large outliers that inflate the true performance of the model; including higher outliers slightly increases the calculated effectiveness of the model.

5.2 Robustness

In the following section the reliability of the results will be addressed. This will be done by evaluating the differences between datasets, the calculated expert scores, with a comparison with the modified version of the Contribution Weighted Method by Budescu and Chen (2014), and the effect of the number of seed questions and experts in an expert group on the performance of the model.

Results per dataset

Figure 11 visualises the results per dataset. For all 45 datasets the average difference from the true value (in terms of standard deviation from the true value per individual question) is shown. The confidence interval lines are used to visualise the intrinsic differences between the datasets. This analysis is used to determine whether the aforementioned results are consistent and affected by outliers. This graph shows a total of three datasets that are significantly different from the rest (or 6.67 percent of the total datasets). These datasets show a significantly lower performance, on average. There are no significantly higher performing datasets. Therefore, the performance is not inflated due to highly performing outliers. Additionally, the results are relatively consistent.

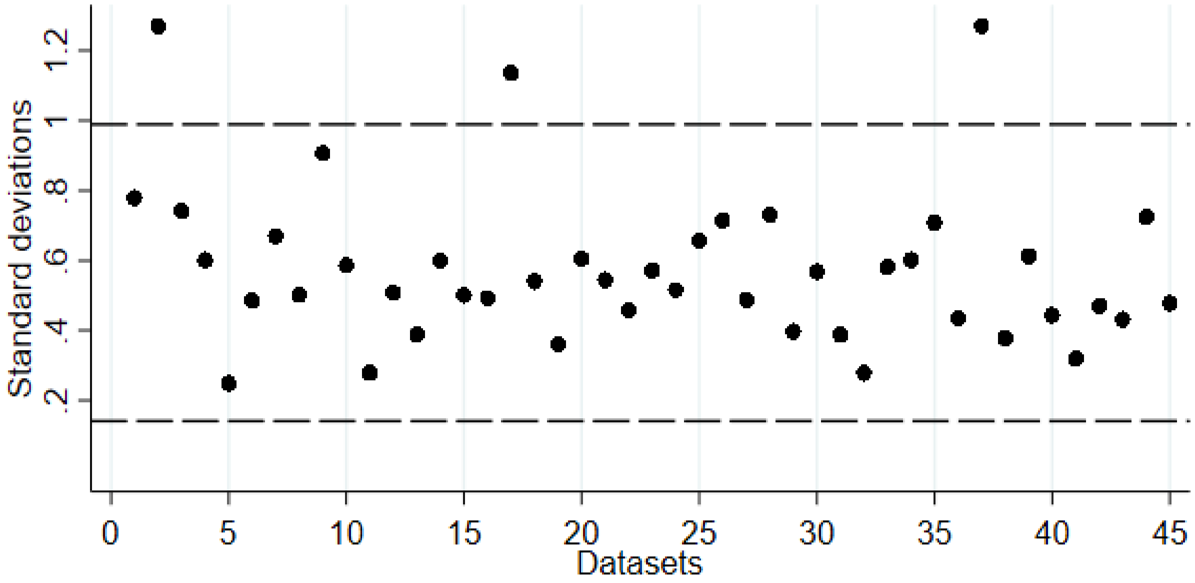


Figure 11. Average standard deviations difference from the true values calculated by the relative accuracy model per dataset (a value closer to zero means the model gives an estimation closer to the true value). The dashed lines depict the upper and lower bound of the 95% confidence interval.

Expert scores

Expert scores are calculated individually for each question in a dataset, then they are combined. Therefore, there is no systematic bias between the calculated expert weight and the accuracy of their prediction. Meaning that the relative accuracy only originates from the expert's capabilities compared to the rest of the group. Therefore, it can be tested whether there is a relationship between the expert scores calculated by the model and the performance of an expert.

A linear regression is used to examine if there is a relationship between the height of the expert scores and the performance of the expert. To prevent bias, similarly to the calculation of the scores, the performance is measured using all questions from a dataset except the question for which the score is calculated. The following linear regression will be used:

$$(12) \qquad Performance = \beta_0 + \beta_1 * \text{Score} + \varepsilon$$

Where *Performance* is the number of standard deviations the estimated value for a seed question differs from the realised value and *Score* the corresponding calculated expert score for a question. β_0 is the constant and ε is the error term. The expert scores are calculated with the relative accuracy model that uses the optimal values for the constants *a* and *b*; these are, respectively, 1 and 3.

The average effect of expert scores on performance using these constants can be seen in table 1. An average significant decrease in difference from the true value with an increase in score is shown. A single point in score increase coincides with an average reduction of 1.66 standard deviations difference from the true value (significant at p<0.1). Indicating that a higher expert score is correlated with better predictions.

Table 1. Results for the linear regression shown in formula (12). Performance in terms of difference from the true value (measured in standard deviations) is the dependent variable. Constants: a=1 and b=3. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

VARIABLES	Performance
Scores	-1.663*
	(0.976)
Constant	2.969***
	(0.934)
Observations	5,640
R-squared	0.001

Comparison with normalised CWM

Different approaches can be takes when calculating weighted expert scores, in terms of intuition. Whereas Budescu and Chen (2014) use the added contribution to the group to measure relative weights in the CWM, the relative accuracy model uses overall accuracy relative to the group to measure relative weights. The intuition from the CWM is used in the normalised CWM where expert scores are calculated using formula (10). This is a modified version of the CWM, where numerical data are used and normalised similarly to the new model introduced in this paper. The relative accuracy method (*accuracy*) and the normalised CWM (*contribution*) are compared in terms of performance in figure 12.

The different approaches are compared using the same dataset. The formula used to compare the two approaches is equation (11); used to measure overall accuracy of the model relative to realised values. The CWM cannot be used on the types of numerical in the TU Delft SEJ dataset since they are not of a binary nature. So, simply the intuition behind the models is compared.

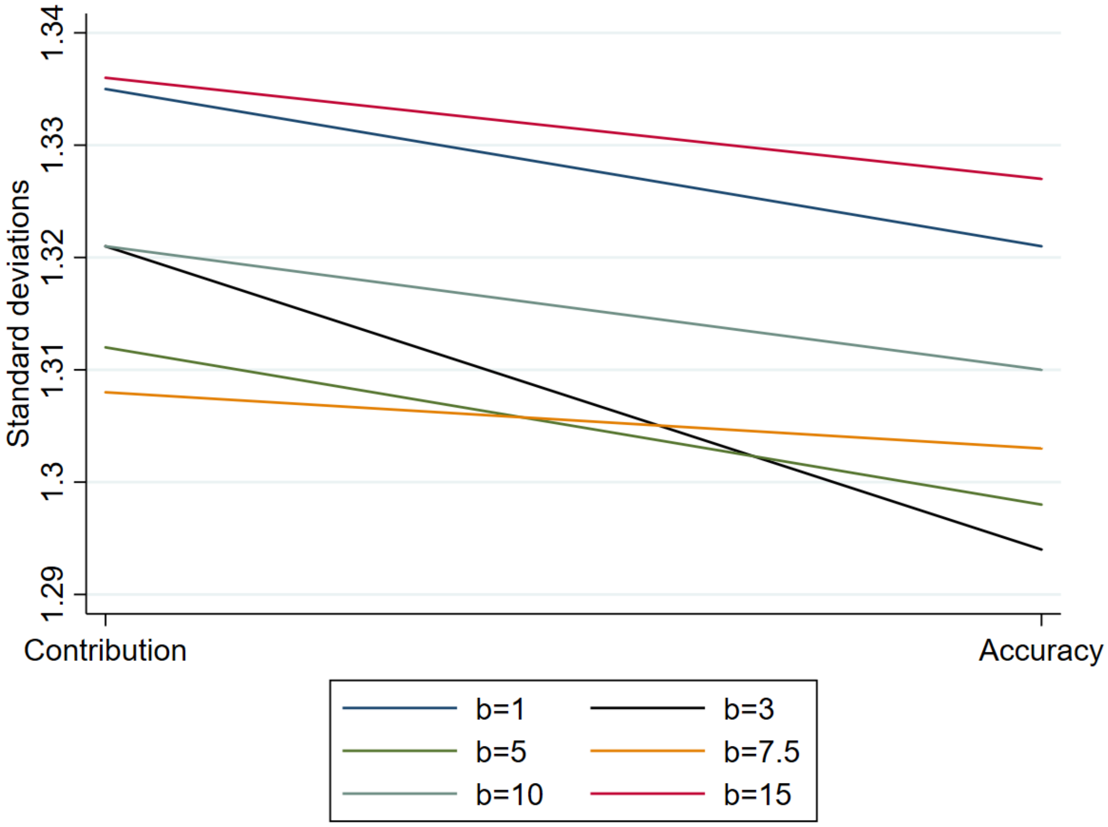


Figure 12. Average performance of the algorithm using the normalised CWM (*contribution*) and the relative accuracy model (*accuracy*). The y-axis shows the difference between the estimated values by the algorithm and the true values in standard deviations. Six different values of *b* are used in the comparison and a=1.

In figure 12 the differences between the two methods are visualised. The accuracy method, on average, has a better performance for the entire range of tested amplification constants (*b*). This is the case when the performance of both models is calculated in terms of the difference in standard deviations with the true value. A more extensive analysis of the performance of the normalised CWM, compared to the best expert and unweighted prediction, for each value of constant *b* is shown in Appendix A. Using the SEJ TU Delft dataset, the normalised CWM method only performs significantly better than the unweighted prediction (Appendix, table 1) and not better than the best expert (Appendix, table 2). Still, this is not the case for the data used in the original paper, where the CWM increases the accuracy of predictions.

Table 2. Differences, in terms of standard deviation away from the true value, between the normalised CWM (*contribution*) and the relative accuracy model (*accuracy*). The differences are calculated for six values of constant *b*. Constant a=1.

b	Contribution	Accuracy	Difference	Difference (p-value)
1	1.335	1.321	0.014	0.000
3	1.321	1.294	0.027	0.001
5	1.312	1.298	0.014	0.001
7.5	1.308	1.303	0.005	0.007
10	1.321	1.310	0.011	0.024
15	1.336	1.327	0.009	0.065

The differences between the two approaches can be seen in table 2. The accuracy method yields a relative lower difference from the true value for all values of *b,* the difference is significant for all values of *b* instead of b=15 (at p<0.05). This indicates that, using the TU Delft SEJ data, predictions by the relative accuracy model significantly outperform predictions by the normalised CWM. Still, no final conclusions can be drawn regarding the absolute difference in performance since a modified version of the CWM was used. Nevertheless, the results indicate that calculating scores based on relative performance slightly improves the final calculations made by the model.

Number of seed questions and experts

Datasets with more seed questions have more information to calculate expert scores with. A linear regression is used to estimate the effect of the total amount of seed questions in a dataset on the performance of the model in that dataset. Clustered standard errors are used since observations within a dataset are expected to be dependent. Where *Performance* is the relative performance in terms of standard deviations away from the true value, β_0 and ε, respectively, the constant and the error term, and *Seed* the number of seed (or calibration) questions in a dataset:

$$(13) \qquad Performance = \beta_0 + \beta_1 * \text{Seed} + \varepsilon$$

A similar clustered linear regression is used to measure the effect of the number of experts in a dataset on the performance in that dataset, where *Experts* is the number of experts in a dataset:

$$(14) \qquad Performance = \beta_0 + \beta_1 * \text{Experts} + \varepsilon$$

Table 3 shows the effects of the amount of seed questions and experts on the performance of the model. In the 45 datasets used, there is no significant increase of performance with an increase in seed questions. Therefore, no conclusions about an optimal amount of seed questions can be made. A higher number of experts does significantly improve the performance of the model. On average, every additional expert significantly decreases the difference between the true value of a question and the estimated value calculated by the model by 0.024 standard deviations (p<0.05).

Table 3. Regression results for the effect of the number of seed questions (13), and the number of experts (14) on the relative performance of the model (a value closer to zero means the model gives an estimation closer to the true value). The linear regressions are shown in formulas (13) and (14). Constants: a=1 and b=3. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

	(13)	(14)
VARIABLES	Performance	Performance
Seed	0.011	
	(0.061)	
Experts		-0.024**
		(0.012)
Constant	1.165*	1.550***
	(0.637)	(0.336)
Observations	494	494
R-squared	0.000	0.001

6. Discussion

The relative accuracy model that is introduced in this paper uses average predictions by experts in datasets where expert data are characterised by its small sample sizes. It calculates expert scores based on how experts perform relative to their peers. Therefore, simple averages in these samples are generally inaccurate due to statistical error. More advanced models are needed to combine expert data to reduce inaccuracy and biases in expert judgement.

The model aims to improve the accuracy of estimations by calculating expert weights based on their relative performance. This means that calculated expert weights increase when an expert has better estimations than the rest of the experts per question. This method therefore does only conform to the classic wisdom of crowds approach to a certain extent. Individual estimations in a usual wisdom of crowds setting can include large outliers of over- and underestimations, which are all included equally in the estimation. This combination of the large variety of data can make remarkably accurate predictions in large samples when no systematic biases exist (Surowiecki, 2005).

Biases also exist in expert judgement. Overconfidence can be used to hide uncertainty (Cesarini, Sandewall, & Johannesson, 2006). Additionally, when experts have made impressive predictions in the past; their confidence increases but their prediction accuracy does not (Denrell & Fang, 2008). Budescu and Chen (2014) use relative contribution to determine expert weights. This is in line with the principal idea of wisdom of crowds. The relative accuracy model, introduced in this paper, uses a different approach. Expert judgements can be combined using all types of numerical data because of the data normalisation method and scores are calculated based on relative performance.

The efficacy of the model is analysed by testing its performance on subjective expert judgement data. This method based on relative performance outperforms the unweighted method and the best expert method. In terms of standard deviations, the model outperforms the unweighted estimation and the best expert by, respectively, 0.05 and 0.15 percent (both significant at $p<0.05$). When comparing the model with the normalised version of the CWM, the relative accuracy model results in more accurate predictions. Additionally, experts with a higher score significantly outperform experts with a lower score. This indicates that expert weights are estimated effectively when calculated by relative performance. It is hard to precisely quantify to what extent the model effectively improves predictions. But, on average, it significantly improves estimations in the TU Delft SEJ dataset. Therefore, the model seems to effectively improve the accuracy of predictions by weighing and combining expert judgements in small samples.

There are, however, some imprecise aspects of the model that need to be addressed. Most importantly, the model uses a cumulative normal distribution to scale the data from zero to one. This is needed to normalise wide varieties of data. In practice, however, a large variety in standard deviations can be observed. This is caused by the data used to calculate expert weights, which are characterized by small sample sizes. When datasets consist of a larger number of experts the performance of the model increases. Whether this is partially caused by an improvement in normalisation or solely by an increase of information to calculate scores with is unclear.

Large outliers, similarly caused by large standard deviations, were excluded when analysing the final performance of the model. This causes the performance tests to lose some of its statistical power. Additionally, the dividing line between outliers and non-outliers in this analysis is chosen arbitrarily; observations were excluded when the standard deviation of a question was more than 50 times larger than the realised value. This caused only extreme outliers to be excluded from the analysis.

The results show that the algorithm improves estimations compared to the unweighted and the best expert measures. But due to the nature of the data it is difficult to quantify precisely how large the improvement is. The relative performance is analysed in terms of standard deviations because the varying data does not allow for a percentual comparison.

The SEJ data from the TU Delft that are used in the testing of the model were originally collected to examine the performance of the classical model (Cooke & Goossens, 2008). This means the data include confidence intervals which are not used in the model. So, data are used which were originally collected with a different purpose. In general, this should not affect results; this can be ensured by testing the model on other data.

Additionally, seed questions were used to determine scores. These questions were designed to measure expertise. The incentives for an expert when answering a seed question could be different from answering a real-world problem within their field of expertise. It is unclear to what extent seed questions are answered differently from real predictions. On the one hand, seed questions could potentially affect risk aversion and confidence of experts. I.e., real predictions are more important because mistakes have larger consequences. Therefore, due to the lack of loss aversion (Rabin & Thaler, 2001), experts could be willing to take more risk when answering seed questions. On the other hand, seed questions are incentivized by feedback; the existence of the true value makes expert predictions verifiable. In certain field of expertise bold predictions are incentivized because rare predictions are perceived as impressive, even though these bold predictions are an indication of lower future accuracy (Denrell & Fang, 2010). An objective calculation of performance could therefore incentivize more precise predictions.

The relative accuracy model slightly outperforms the normalised CWM, which is a modified version of the CWM (Budescu & Chen, 2014), using the SEJ TU Delft data. This indicates that the method of calculating scores is improved by using relative accuracy. Since the models use different types of data no straightforward comparison can be made, however.

Additionally, there is some proof that weighted expert models do not perform in out of sample data, which is the case for the Classical Model (Clemen, 2008). In practice, however, weighted models seem to outperform simple methods such as taking the average (Lin & Cheng, 2009). Indicating that Cooke's model can still be a useful way to make predictions, but that the true usefulness may have been overstated. How the method introduced in this paper performs with other data is still uncertain.

7. Conclusion

The wisdom of crowds, combining a large number of estimations, is generally an effective way of estimating answers to real-life questions. However, when sample sizes are too small to reduce statistical bias other approaches are needed. There are different methods to improve predictions due to this uncertainty. Currently, the CM (Cooke & Goossens, 2008) and the CWM (Budescu & Chen, 2014) are models that can calculate an aggregate estimation using small sample sizes. Alternative behavioural methods to elicit combined expert predictions exist, such as consultations between experts (Rohrbaugh, 1979, 1981; Flores & White, 1989). This paper introduces the relative performance model which calculates estimations based on expert weights determined by their relative accuracy. This model, alternatively, makes use of a more practical approach than current models since it can be used with any type of numerical data.

The combined predictions by the relative accuracy model outperform both unweighted predictions and the best expert per expert group. Therefore, expert bias is effectively reduced by weighing and combining individual predictions. It is still unclear what improvements can currently be made and how the model compares to its alternatives in practice. It is difficult to test performance between models since different types of data are used. The advantage of the relative accuracy model is that no expert predictions that include confidence intervals are needed, but only a single prediction per expert. Predicted confidence intervals can be affected by biases, such as overconfidence (Soll & Klayman 2004). Additionally, since no confidence intervals are needed, the model can more easily be applied in real-world cases.

Therefore, the model could be tested by collecting experts' past predictions and evaluating how effective future predictions are. Or, more generally, by testing the model on different datasets. Furthermore, this model was tested to be used in small (expert) samples. It could be examined whether the model also outperforms a wisdom of crowds approach in larger samples. These suggestions for future research are essential for determining whether the model performs consistently.

All in all, the results indicate that the relative accuracy model based on relative performance can improve predictions. Additional research is needed to determine if the model can be applied in practice.

References

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society,* 20(4), 451-468.

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

Cesarini, D., Sandewall, Ö., & Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior & Organization,* 61(3), 453-470.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5), 1367-1403.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4), 559-583.

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety*, 93(5), 760-765.

Cooke, R. M., & Goossens, L. L. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5), 657-674.

Denrell, J., & Fang, C. (2010). Predicting the next big thing: Success as a signal of poor judgment. *Management Science*, 56(10), 1653-1667.

Evgeniou, T., Fang, L., Hogarth, R. H., & Karelaia, N. (2013). Competitive dynamics in forecasting: The interaction of skill and uncertainty. *Journal of Behavioral Decision Making* 26(4), 375–384.

Flores, B. E., & White, E. M. (1989). Subjective versus objective combining of forecasts: an experiment. *Journal of Forecasting*, 8(3), 331-341.

Fischhoff, B., & Beyth, R. (1975). I knew it would happen: Remembered probabilities of once-future things. *Organizational Behavior and Human Performance*, 13(1), 1-16.

Galton, F. (1907). Vox Populi. *Nature,* 75, 450-451.

Hilary, G., & Menzly, L. (2006). Does past success lead analysts to become overconfident? *Management science*, 52(4), 489-500.

Kahneman, D., Slovic, P., Tversky, A., eds (1982) Judgment Under Uncertainty: Heuristics and Biases (Cambridge University Press, New York).

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6), 1121-1134.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, 52(1), 111-127.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition*, 39(5), 914-923.

Lin, S. W., & Cheng, C. H. (2009). The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management*, 4(2), 149-161.

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.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, 111(20), 7176-7184.

Rabin, M., & Thaler, R. H. (2001). Anomalies: risk aversion. *Journal of Economic perspectives*, 15(1), 219-232.

Rohrbaugh, J. (1979). Improving the quality of group judgment: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance*, 24(1), 73-92.

Rohrbaugh, J. (1981). Improving the quality of group judgment: Social judgment analysis and the nominal group technique. *Organizational Behavior and Human Performance*, 28(2), 272-288.

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

Soll, J. B., & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 299–314.

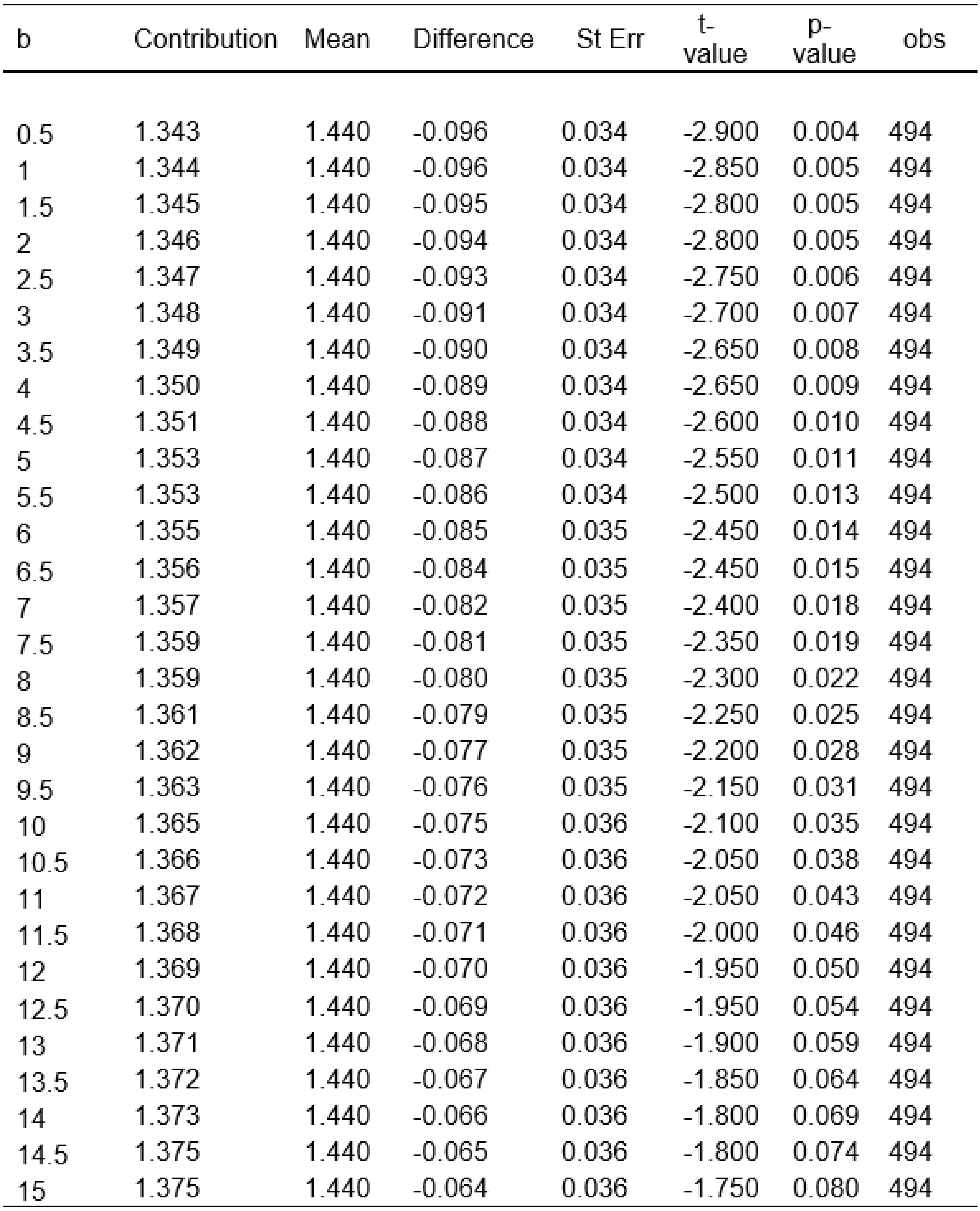
Surowiecki, J. (2005). The wisdom of crowds. Anchor.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: *Heuristics and biases. science*, 185(4157), 1124-1131.

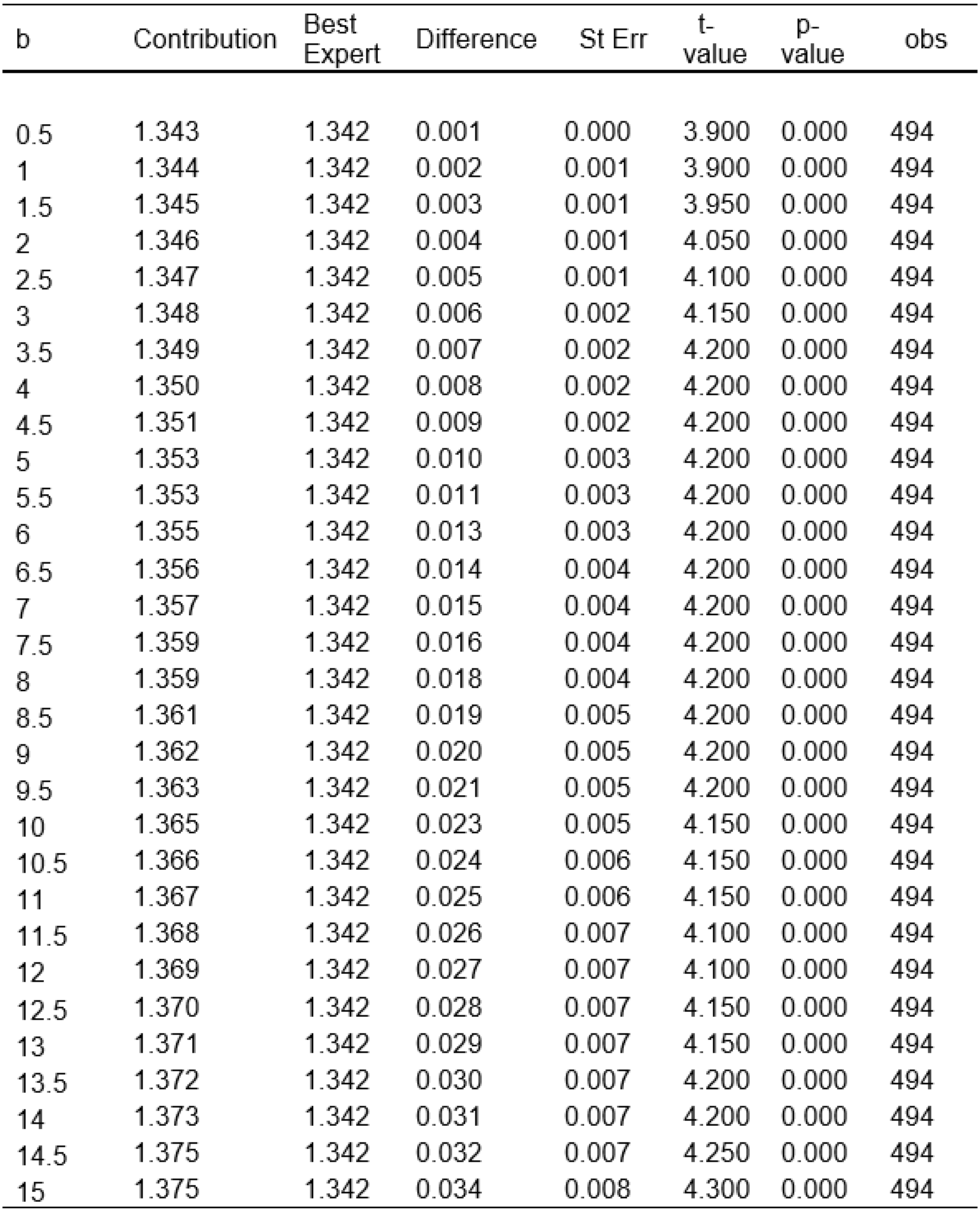
Appendix

Table 1. Difference in performance of the normalised CWM (*contribution*) and the unweighted score (*mean*). Performance is calculated in standard deviations difference from the true value. The performance is calculated for different values of constant *b*, ranging for 0.5 to 15. Constant a=1.



b	Contribution	Mean	Difference	St Err	t-value	p-value	obs
0.5	1.343	1.440	-0.096	0.034	-2.900	0.004	494
1	1.344	1.440	-0.096	0.034	-2.850	0.005	494
1.5	1.345	1.440	-0.095	0.034	-2.800	0.005	494
2	1.346	1.440	-0.094	0.034	-2.800	0.005	494
2.5	1.347	1.440	-0.093	0.034	-2.750	0.006	494
3	1.348	1.440	-0.091	0.034	-2.700	0.007	494
3.5	1.349	1.440	-0.090	0.034	-2.650	0.008	494
4	1.350	1.440	-0.089	0.034	-2.650	0.009	494
4.5	1.351	1.440	-0.088	0.034	-2.600	0.010	494
5	1.353	1.440	-0.087	0.034	-2.550	0.011	494
5.5	1.353	1.440	-0.086	0.034	-2.500	0.013	494
6	1.355	1.440	-0.085	0.035	-2.450	0.014	494
6.5	1.356	1.440	-0.084	0.035	-2.450	0.015	494
7	1.357	1.440	-0.082	0.035	-2.400	0.018	494
7.5	1.359	1.440	-0.081	0.035	-2.350	0.019	494
8	1.359	1.440	-0.080	0.035	-2.300	0.022	494
8.5	1.361	1.440	-0.079	0.035	-2.250	0.025	494
9	1.362	1.440	-0.077	0.035	-2.200	0.028	494
9.5	1.363	1.440	-0.076	0.035	-2.150	0.031	494
10	1.365	1.440	-0.075	0.036	-2.100	0.035	494
10.5	1.366	1.440	-0.073	0.036	-2.050	0.038	494
11	1.367	1.440	-0.072	0.036	-2.050	0.043	494
11.5	1.368	1.440	-0.071	0.036	-2.000	0.046	494
12	1.369	1.440	-0.070	0.036	-1.950	0.050	494
12.5	1.370	1.440	-0.069	0.036	-1.950	0.054	494
13	1.371	1.440	-0.068	0.036	-1.900	0.059	494
13.5	1.372	1.440	-0.067	0.036	-1.850	0.064	494
14	1.373	1.440	-0.066	0.036	-1.800	0.069	494
14.5	1.375	1.440	-0.065	0.036	-1.800	0.074	494
15	1.375	1.440	-0.064	0.036	-1.750	0.080	494

Table 2. Difference in performance of the normalised CWM (*contribution*) and the Best Expert. Performance is calculated in standard deviations difference from the true value. The performance is calculated for different values of constant *b*, ranging for 0.5 to 15. Constant a=1.



b	Contribution	Best Expert	Difference	St Err	t-value	p-value	obs
0.5	1.343	1.342	0.001	0.000	3.900	0.000	494
1	1.344	1.342	0.002	0.001	3.900	0.000	494
1.5	1.345	1.342	0.003	0.001	3.950	0.000	494
2	1.346	1.342	0.004	0.001	4.050	0.000	494
2.5	1.347	1.342	0.005	0.001	4.100	0.000	494
3	1.348	1.342	0.006	0.002	4.150	0.000	494
3.5	1.349	1.342	0.007	0.002	4.200	0.000	494
4	1.350	1.342	0.008	0.002	4.200	0.000	494
4.5	1.351	1.342	0.009	0.002	4.200	0.000	494
5	1.353	1.342	0.010	0.003	4.200	0.000	494
5.5	1.353	1.342	0.011	0.003	4.200	0.000	494
6	1.355	1.342	0.013	0.003	4.200	0.000	494
6.5	1.356	1.342	0.014	0.004	4.200	0.000	494
7	1.357	1.342	0.015	0.004	4.200	0.000	494
7.5	1.359	1.342	0.016	0.004	4.200	0.000	494
8	1.359	1.342	0.018	0.004	4.200	0.000	494
8.5	1.361	1.342	0.019	0.005	4.200	0.000	494
9	1.362	1.342	0.020	0.005	4.200	0.000	494
9.5	1.363	1.342	0.021	0.005	4.200	0.000	494
10	1.365	1.342	0.023	0.005	4.150	0.000	494
10.5	1.366	1.342	0.024	0.006	4.150	0.000	494
11	1.367	1.342	0.025	0.006	4.150	0.000	494
11.5	1.368	1.342	0.026	0.007	4.100	0.000	494
12	1.369	1.342	0.027	0.007	4.100	0.000	494
12.5	1.370	1.342	0.028	0.007	4.150	0.000	494
13	1.371	1.342	0.029	0.007	4.150	0.000	494
13.5	1.372	1.342	0.030	0.007	4.200	0.000	494
14	1.373	1.342	0.031	0.007	4.200	0.000	494
14.5	1.375	1.342	0.032	0.007	4.250	0.000	494
15	1.375	1.342	0.034	0.008	4.300	0.000	494