

Contemporary methods for identifying and leveraging
expertise in collective decision-making

Marcellin Ferdinand Ruxuan Carlos Martinie

ORCID 0000-0002-1289-1467

Submitted in total fulfilment of the requirements of the
degree of Doctor of Philosophy

Melbourne School of Psychological Sciences
The University of Melbourne

May 2020

Abstract

From company board committees to grand juries to political parties, decision making often involves a group of individuals with differing views on what the best decision may be. Individuals have different views, in part, because some individuals have more knowledge than others. In theory, decision makers can make better decisions by leveraging the expertise of individuals in the crowd by identifying those who have more knowledge than others. Unfortunately, it is often difficult to identify and leverage expertise in practice. Decision makers often have no records of individuals' past performance from which they can estimate expertise, and subjective measures such as confidence and self-ratings of expertise are often considered as unreliable. In this thesis, I demonstrate how individuals' meta-predictions – predictions about what other individuals will predict – can be used to identify and leverage expertise in the crowd on 'single-question' problems, where records on individuals' past performance are unavailable. I first examine how a recent algorithm in the literature can be used to distinguish between subsets of high-performing and low-performing individuals in the crowd on binary decision problems. I show that this algorithm is in fact weighting individuals by the absolute difference between an individual's vote and their meta-prediction about the votes of others, and thus this weighting metric can be used to identify and leverage expertise in the crowd. I develop an improved weighting approach that uses individuals' probabilistic forecasts and their meta-predictions about the average probability forecast of others, and show that this outperforms the top alternative probabilistic aggregation approaches in the literature on a large range of decision problems. Furthermore, I demonstrate that this improved weighting approach provides a superior measure of expertise than existing single-question approaches to identifying expertise in the literature. As an additional test, I compare this single-question approach with cross-domain weighting – weighting by individuals' performance on problems on unrelated domains – and show that cross-domain weighting is favoured over single-question approaches in cases where it is possible to obtain individuals' past performance on problems – even if those questions are from unrelated domains. In general, our results demonstrate the potential for using individuals' meta-predictions and performance on problems from unrelated domains to identify expertise in cases where other approaches might be ineffective or unavailable.

Declaration

I declare that: (1) the thesis comprises only original work towards a degree of Doctor of Philosophy in the Faculty of Medicine, Dentistry and Health Sciences (Psychology), except where indicated; (2) due acknowledgement has been made in the text to all other material used; and (3) the thesis contains fewer words than the maximum word limit, exclusive of tables, maps, bibliographies and appendices.

Marcellin Ferdinand Ruxuan Carlos Martinie

Preface

This thesis is a compilation of original work by Marcellin Martinie. Chapters 2–4 were written as thesis chapters and subsequently adapted, in part, for submission as manuscripts to peer-reviewed journals. Specifically:

- Chapter 2. Some of the empirical analyses in this chapter were adapted for the manuscript, Wilkening, Martinie, and Howe (2020), which is currently under revision following peer review in *Management Science*.
- Chapter 3. The core ideas and findings in this chapter have been adapted for the paper, Martinie, Wilkening, and Howe (2020), which was published in *PLOS One* on 24 April 2020.
- Chapter 4. The manuscript based on this chapter is currently in preparation.

Data collection, data analysis, and written drafts were completed entirely by Marcellin Martinie for the work included in the main text of this thesis. Drafts were revised and edited by Marcellin Martinie and his supervisors, A/Prof. Piers Howe, and Prof. Tom Wilkening.

Research in this thesis was supported by the Australian Government’s Research Training Program Scholarship to Marcellin Martinie, the FBE & MDHS Collaboration Seed Funding Award to A/Prof. Piers Howe and Prof. Tom Wilkening, and the Australian Research Council’s Discovery Early Career Research Award DE140101014 to Prof. Tom Wilkening.

Acknowledgements

I have met many wonderful people over the course of my PhD, and it brings me joy to be able to acknowledge how much they have helped me on this journey.

To my supervisors, Piers Howe and Tom Wilkening, thank you for your guidance, counsel, and patience over the years. Your teachings have been invaluable, and I could not have asked for better mentors.

To my dear friends: Xian, Sarah, Jun, Weijia, Jess, and Campbell, thank you for enriching my life with happiness and wonder, for always being there for me, and for entertaining my shenanigans over the years.

To my friends and colleagues from Melbourne University: Larson, Kathryn, Annie, Dave, Ariel, Paul, Daryl, Lauren, Deb, Geoff, Maggie, Simon, Jason, Gabe, and Felix, thank you for being such amazing people and for making this journey as memorable as it was.

To my family: Mum, Lysh, Dani, and Dan, thank you for your overwhelming, unconditional, and endless support over the years. This journey would not have been possible without you. It is to you that I dedicate this thesis.

Contents

1	Introduction	1
1.1	The Wisdom of Crowds	2
1.2	Weighted Aggregation Approaches	3
1.3	Cognitive Modeling Approaches	6
1.4	Single-Question Aggregation Approaches	7
1.5	Outline of the Present Research	9
2	Leveraging expertise using crowd meta-predictions	11
2.1	Introduction	12
2.1.1	Weighting by past performance	13
2.1.2	Confidence is an unreliable predictor of expertise	14
2.1.3	Differences in meta-cognitive knowledge between experts and novices . . .	15
2.1.4	The Surprisingly Popular algorithm	16
2.1.5	Questions to be addressed	19
2.2	Preliminaries	20
2.2.1	Individual responses and predictions	20
2.2.2	Measures of binary classification performance	20
2.2.3	The Surprisingly Popular algorithm	22
2.2.4	Other categorical forecast-aggregation algorithms	25
2.3	A theoretical model for understanding the SP mechanism	28
2.4	Experiment 1	32

2.4.1	Methods	32
2.4.2	Results and Discussion	35
2.4.3	Analysis of expertise	38
2.4.4	Identifying the SP weights assigned to individuals	42
2.4.5	Distribution of forecasters' average accuracy	44
2.4.6	Simulating changes in expertise	44
2.5	Experiment 2	49
2.5.1	Methods	49
2.5.2	Results	50
2.5.3	Analysis of expertise	54
2.5.4	Weights assigned by the SP algorithm	55
2.5.5	Distribution of forecasters' percentage accuracy	58
2.5.6	Discussion	60
2.6	Robustness Over Different Sample Sizes	61
2.7	Testing the SP Mechanism on NFL Predictions	63
2.8	General Discussion	67
2.8.1	Contributions of the present research	67
2.8.2	Connection to existing research on the SP algorithm	69
2.8.3	Connection to other expertise-identification approaches	69
2.8.4	Considerations for future research	70
3	Probabilistic single-question forecasting approaches	72
3.1	Probabilistic Forecasting	72
3.1.1	Quantifying uncertainty	72
3.1.2	The goal of probabilistic forecasting	74
3.1.3	Scoring rules	74
3.1.4	The Wisdom of Crowds literature	75
3.1.5	Weighted forecast-aggregation models	78
3.1.6	Recalibration and extremisation approaches to forecast aggregation	79
3.1.7	Single-question aggregation approaches	82

3.2	Experiment 1: Testing categorical forecast-aggregation approaches in the probabilistic forecasting domain	88
3.2.1	Methods	89
3.2.2	Individuals' responses	91
3.2.3	Algorithms	91
3.2.4	Results and Discussion	95
3.3	The Meta-Vote Weighting Algorithm	100
3.4	Experiment 2: Validating the Meta-Vote Weighting Algorithm	101
3.4.1	Methods	103
3.4.2	Results	103
3.4.3	Optimally recalibrating model predictions	109
3.4.4	Discussion	111
3.5	Experiment 3: A Refined Measure of Crowd Expertise	116
3.5.1	Methods	118
3.5.2	Results	119
3.5.3	Discussion	123
3.6	Experiment 4: Validating the Meta-Probability Weighting Algorithm	124
3.6.1	Methods	124
3.6.2	Analyses	125
3.6.3	Results	127
3.6.4	Discussion	132
3.6.5	Supplementary analyses	133
3.7	General Discussion	138
3.7.1	Contributions of the present research	141
3.7.2	Relationship to previous research	142
3.7.3	Considerations for future research	143
4	Identifying expertise via cross-domain performance	145
4.1	Abstract	145
4.2	Introduction	146

4.3	Research Design	150
4.4	Experiment 1: Comparing cross-domain weighting to within-domain weighting and single-question aggregation approaches	153
4.4.1	Methods	154
4.4.2	Analyses	154
4.4.3	Results	158
4.4.4	Discussion	159
4.5	Experiment 2: Testing cross-domain weighting using performance on multiple unrelated domains	160
4.5.1	Methods	161
4.5.2	Analyses	162
4.5.3	Results	163
4.5.4	Discussion	164
4.6	Experiment 3: Art, Science Trivia, and Emotional Intelligence problems	167
4.6.1	Methods	168
4.6.2	Analyses	168
4.6.3	Results	169
4.6.4	Discussion	169
4.7	Experiment 4: Replicating Experiment 3's results using a larger sample	171
4.7.1	Methods	172
4.7.2	Analyses	173
4.7.3	Results	173
4.7.4	Discussion	174
4.7.5	Post-hoc Simulations	176
4.7.6	Comparing cross-domain weighting to other single-question aggregation approaches	179
4.8	General Discussion	179
4.8.1	Conclusions	183
5	Concluding Remarks	184

6	Appendices	196
6.1	Chapter 2 Appendices	196
6.1.1	Manuscript detailing the theoretical model proposed in Chapter 2	196
6.1.2	List of questions from Experiment 1: US States dataset from Chapter 2 .	263
6.1.3	List of questions from Experiment 2: US Grades dataset from Chapter 2 .	266
6.2	Chapter 3 Appendices	279
6.2.1	Manuscript adapted from Chapter 3	279
6.3	Chapter 4 Appendices	311
6.3.1	List of questions from each experiment in Chapter 4	311

List of Tables

2.1	Formula and Descriptions of Each Aggregation Algorithm. See Section 2.2.1 for Details on Notation.	27
2.2	Mean Difference in Cohen’s Kappa Coefficient between the SP algorithm and Each Other Algorithm for Each of Prelec, Seung, and McCoy’s (2017) Datasets and Our Experiment 1 Dataset	37
2.3	Mean Difference in Cohen’s Kappa Coefficient Between the SP Algorithm and Each Other Algorithm for Each Dataset from Experiment 2	53
3.1	Details for each dataset used in this chapter	90
3.2	Binary aggregation approaches	93
3.3	Probabilistic aggregation approaches	94

List of Figures

2.1	Example of a trial in Experiment 1.	34
2.2	Classification performance of algorithms measured by Cohen’s Kappa Coefficient on each of Prelec et al.’s (2017) five datasets (left), and our dataset for Experiment 1 (right). Error bars show standard error.	36
2.3	Classification performance of algorithms measured by percentage accuracy on each of Prelec et al.’s (2017) five datasets (left), and our dataset for Experiment 1 (right). Error bars show the standard error.	36
2.4	Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the States (MWH) dataset. Each point represents that group’s average vote and meta-prediction for one event in the dataset. The diagonal line indicates where that group’s meta-predictions are exactly equal to their proportion of votes for “true”. The shaded regions indicate where the SP algorithm would generate correct predictions.	39
2.5	Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the States (PSM) dataset.	40
2.6	Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Trivia dataset.	40

2.7	Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Lesions dataset.	41
2.8	Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Art (N) dataset.	41
2.9	Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Art (E) dataset.	42
2.10	The average weight assigned by the SP algorithm as a function of forecasters' accuracy in each of the six datasets analysed in Experiment 1	43
2.11	The distribution of forecasters over the percentage of events correctly predicted from our US States dataset from Experiment 1	45
2.12	The change in Cohen's Kappa Coefficient of each algorithm as expertise is manipulated	47
2.13	The change in percentage accuracy of each algorithm as expertise is manipulated	48
2.14	Example of a trial in Experiment 2.	51
2.15	The mean Cohen's Kappa Coefficient and standard error for each algorithm across each level of question difficulty (1 - easiest to 5 - hardest).	52
2.16	The mean percentage accuracy and standard error for each algorithm across each level of question difficulty (1 - easiest to 5 - hardest).	52
2.17	Proportion of high-performing individuals and low-performing individuals voting true relative to their meta-predictions for questions where the outcome was True (left) and False (right) for our Experiment 2 US Grades Difficulty 1 dataset . . .	55
2.18	Proportion of high-performing individuals and low-performing individuals voting true relative to their meta-predictions for questions where the outcome was True (left) and False (right) for our Experiment 2 US Grades Difficulty 2 dataset . . .	56
2.19	Proportion of high-performing individuals and low-performing individuals voting true relative to their meta-predictions for questions where the outcome was True (left) and False (right) for our Experiment 2 US Grades Difficulty 3 dataset . . .	56

2.20	Proportion of high-performing individuals and low-performing individuals voting true relative to their meta-predictions for questions where the outcome was True (left) and False (right) for our Experiment 2 US Grades Difficulty 4 dataset	57
2.21	Proportion of high-performing individuals and low-performing individuals voting true relative to their meta-predictions for questions where the outcome was True (left) and False (right) for our Experiment 2 US Grades Difficulty 5 dataset	57
2.22	The average weight assigned by the SP algorithm as a function of forecasters' accuracy in each of the five US Grades datasets in Experiment 2	59
2.23	The distribution of forecasters over the percentage of events correctly predicted in each of the five datasets from Experiment 2	60
2.24	Simulation results showing the change in algorithms' performance in terms of Cohen's Kappa Coefficient over different sample sizes for each of the five datasets from Experiment 2. Error bars show the standard error.	62
2.25	Percentage accuracy for each algorithm for each week from Lee, Danileiko, and Vi's (2018) NFL dataset.	64
2.26	Proportion of low-performing individuals (blue crosses) and high-performing individuals (red circles) voting true relative to their meta-predictions for questions from Lee et al.'s (2018) dataset	65
2.27	The average weight assigned by the SP algorithm to the votes of forecasters as a function of the performance of the forecasters for the events in Lee et al.'s (2018)NFL dataset	66
2.28	The distribution of forecasters over the percentage of events correctly predicted from each week in Lee et al.'s (2018)NFL dataset	67
3.1	The mean transformed Brier score for the Majority Vote (MV), Confidence-weighted (CW), Max confidence (MC), and Surprisingly popular algorithm (SP), the unweighted mean (UM), Recalibrated unweighted mean (UM-R), the P''_{cs} aggregator (P''_{cs}), and the Minimal Pivoting (MP) algorithm, on 1196 events across 12 datasets. Error bars show the standard error.	96

3.2	The calibration curve for the unweighted mean (orange), the Recalibrated unweighted mean (red), the P''_{cs} aggregator (light green), and the Minimal Pivoting model (blue). Each calibration curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the calibration of a perfectly-calibrated forecaster (or algorithm) whose forecasted probabilities match exactly the proportion of “true” events observed.	98
3.3	The mean transformed Brier score for the Surprisingly Popular (SP) unweighted mean (UM), P''_{cs} , Minimal Pivoting (MP), and Meta-Vote Weighting (MVW) algorithms over 1196 events across 12 datasets (light grey), along with the recalibrated version of each algorithm (dark grey). Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms, recalibrated or otherwise.	104
3.4	The calibration curve for the unweighted mean (orange), the P''_{cs} aggregator (red), the Minimal Pivoting model (light green), the Meta-Vote Weighting algorithm (light blue), and the Recalibrated Meta-Vote Weighting algorithm (dark blue). Each calibration curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the calibration of a perfectly-calibrated forecaster (or algorithm), whose forecasted probabilities match exactly the proportion of “true” events observed.	107
3.5	The mean transformed Brier score for the Surprisingly Popular algorithm (yellow), the unweighted mean (orange), the P''_{cs} aggregator (red), the Minimal Pivoting model (light green), the Meta-Vote Weighting algorithm (light blue), and the Recalibrated Meta-Vote Weighting algorithm (dark blue) on each individual dataset. Error bars show the standard error.	108

3.6	The mean transformed Brier score for the optimally recalibrated versions of the unweighted mean (UM), the P''_{cs} aggregator, and the Minimal Pivoting (MP) algorithm, compared to the Recalibrated Meta-Vote Weighting (MVW) algorithm, which has not been optimally recalibrated. Error bars show the standard error. The Recalibrated Meta-Vote Weighting algorithm (far right) still significantly outperforms all other algorithms.	112
3.7	The mean transformed Brier score for the unweighted mean (UM), the P''_{cs} aggregator, the Minimal Pivoting (MP) algorithm, the Meta-Vote Weighting (MVW) algorithm, the Recalibrated Meta-Vote Weighting (MVW-R) algorithm, the Meta-Probability Weighting algorithm (MPW), and the Recalibrated Meta-Probability Weighting (MPW-R) algorithm over 1196 events across 12 datasets. Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms.	119
3.8	The calibration curve for the Meta-Vote Weighting algorithm (orange), the Recalibrated Meta-Vote Weighting algorithm (red), the Meta-Probability Weighting algorithm (light green), and the Recalibrated Meta-Probability Weighting algorithm (blue). Each calibration curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the curve of a perfectly calibrated forecaster (or algorithm), whose forecasted probabilities match exactly the proportion of “true” events observed.	121
3.9	The mean transformed Brier score for the unweighted mean (yellow), the P''_{cs} aggregator (orange), the Minimal Pivoting model (red), the Meta-Vote Weighting algorithm (light green), the Recalibrated Meta-Vote Weighting algorithm (light blue), the Meta-Probability Weighting algorithm (dark blue), and the Recalibrated Meta-Probability Weighting algorithm (purple) on each individual dataset. Error bars show the standard error.	122
3.10	Example of a trial in Experiment 4 – replication of the US Grades experiment. .	126

3.11	The mean transformed Brier score for the unweighted mean (UM), the P''_{cs} aggregator, the Minimal Pivoting (MP) algorithm, the Meta-Vote Weighting (MVW) algorithm, the Recalibrated Meta-Vote Weighting (MVW-R) algorithm, the Meta-Probability Weighting algorithm (MPW), and the Recalibrated Meta-Probability Weighting (MPW-R) algorithm over 500 US Grades questions varying across five levels of difficulty. Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms.	128
3.12	The calibration curve for the Meta-Vote Weighting algorithm (orange), the Recalibrated Meta-Vote Weighting algorithm (red), the Meta-Probability Weighting algorithm (light green), and the Recalibrated Meta-Probability Weighting algorithm (blue). Each curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the calibration of a perfectly-calibrated forecaster (or algorithm), whose forecasted probabilities match exactly the proportion of “true” events observed.	130
3.13	The mean transformed Brier score for the unweighted mean (yellow), the P''_{cs} aggregator (orange), the Minimal Pivoting model (red), the Meta-Vote Weighting algorithm (light green), the Recalibrated Meta-Vote Weighting algorithm (light blue), the Meta-Probability Weighting algorithm (dark blue), and the Recalibrated Meta-Probability Weighting algorithm (purple) on each US Grades dataset, in order of increasing difficulty from lowest difficulty level (Difficulty 1) to highest difficulty level (Difficulty 5). Error bars show the standard error.	131
3.14	The mean transformed Brier score for the optimally recalibrated versions of the unweighted mean (UM), the P''_{cs} aggregator, and the Minimal Pivoting (MP) algorithm, compared to the Meta-Probability Weighting (MPW) algorithm, which has not been optimally recalibrated. Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms.	132

3.15	Simulation results showing the change in score for the unweighted mean (UM), Meta-Vote Weighting (MVW) algorithm, Recalibrated Meta-Vote Weighting (MVW-R) algorithm, Meta-Probability Weighting (MPW) algorithm, and Recalibrated Meta-Probability Weighting (MPW-R) algorithm over different sample sizes for each level of question difficulty in Experiment 4. Error bars show the standard error. .	134
3.16	The average weight assigned by the Meta-Vote Weighting (and SP) algorithm, the Meta-Probability Weighting algorithm, and the Decision Similarity weighting model to the votes of forecasters as a function of the performance of the forecasters for each of the five difficulties and overall across all five difficulties. Forecasters were ranked by performance in terms of percentage accuracy on each event. Error bars represent the standard error.	137
3.17	Performance for each of the weighting metrics after averaging over the forecasts of the top n percentage of forecasters selected from the training set using leave-one-out cross-validation. Forecasters were selected based on their average weight in the training set, determined by either their Meta-Probability weights, Meta-Vote weights, or their Decision Similarity. Error bars represent the standard error. Meta-Probability Weights appear to consistently outperform both other weighting metrics.	139
4.1	The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting algorithm, the Recalibrated Meta-Probability Weighting algorithm, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the NFL Trivia (left) and Science Trivia domain (right). Error bars show the standard error. The CWM significantly outperformed the xCWM in the NFL Trivia domain but not in the Science Trivia domain. . . .	159

4.2	The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the NFL Trivia domain (left) and Science Trivia domain (right). Error bars show the standard error. There was no significant difference between the performance of the CWM and the Emotional Intelligence (EI) weights in either domain.	165
4.3	The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM), NFL Weights, Science Weights, and the cross-domain Contribution-Weighted Model (xCWM) on questions from the Emotional Intelligence (EI) domain. Error bars show the standard error. There was no significant difference between the performance of the CWM and NFL Weights, Science Weights, or the xCWM. . .	166
4.4	The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence questions (right). Error bars show the standard error. The xCWM significantly outperformed the CWM in the Science Trivia and Emotional Intelligence domains, but not in the Art domain.	170
4.5	The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence (EI) questions (right). Error bars show the standard error. There was no significant difference between the performance of the xCWM and CWM in any of the three domains. .	175

4.6	Simulation using results from Experiment 1-4 showing the mean transformed Brier score for the Contribution-Weighted Model (red), which is trained on questions from the same domain as the test domain, compared to the cross-domain Contribution-Weighted Model (xCWM), which is trained on questions from a different domain to the test domain. Performance is calculated across all questions in each experiment, shown as a function of training set size. The performance of the simple average (dashed line), which does not use training data, is shown for reference.	178
4.7	The mean transformed Brier score for the Unweighted Mean, Decision Similarity Weighting, the Minimal Pivoting approach, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence (EI) questions (right). Error bars show the standard error.	180

1 | Introduction

Decision making is fundamental to everyday life. In order to make decisions, we need to gather and integrate relevant information from our internal and external environment. The notion that we can improve decision making by combining the decisions from other people has been termed the *Wisdom of Crowds* – a phenomenon which has been shown to be widely robust across a diverse range of decision and forecasting problems. Of notable interest has been the use of crowd wisdom in predicting the outcomes of political elections, predicting the change in unemployment and inflation rate from year to year in Europe, and the classification of a range of medical diagnoses (Budescu & Chen, 2015; Prelec et al., 2017; Surowiecki, 2005). While there is strong scientific consensus that judgments can be improved by aggregation, the choice and application of different aggregation approaches remains an active area of research in the literature. Different aggregation approaches are suited to different types of problems and unsurprisingly, there is no simple one-size-fits-all solution to the forecast-aggregation problem. Despite much progress being made recently in the development of better aggregation approaches, there remain many unaddressed questions in the literature, particularly regarding approaches that have been developed in the last five years. The goal of this thesis is to provide insight into the efficacy and limitations of some of these recent aggregation approaches and to develop approaches that better address these limitations.

In this chapter, we will provide a review of the relevant literature, introduce the issues in the literature that the later chapters in this thesis will aim to address, and outline the structure for the remaining chapters in this thesis.

1.1 The Wisdom of Crowds

The idea that combining information from multiple people can produce better predictions or decision outcomes has been well-known for centuries. In political science, the Jury theorem proposed by Condorcet (1785) proves that the probability of majority voting producing the correct decision increases monotonically with the number of voters, under the assumption that the average probability individuals voting for the correct decision is higher than that for the incorrect decision. While the Condorcet Jury Theorem is only directly relevant to binary decisions, its principles have since been extended to probabilistic forecasting (Murr, 2015).

Over a century after Condorcet’s Jury Theorem was proposed, Galton (1907) provided one of the first rigorous demonstrations of crowd wisdom in quantity estimation in which he compared the error associated with individual and crowd estimates of the weight of a butchered and dressed ox at a livestock exhibition. Galton found that while most individual forecasters estimated quite far from the true weight, the mean and median forecasts of the crowd were extremely accurate.

In the subsequent decades, several other authors had reported findings consistent with crowd wisdom in estimation and prediction tasks. In perhaps the earliest work after Galton’s, Knight (1921) reported several experiments on crowd wisdom, including one where participants rank ordered the IQ of children after looking at their photographs and one where participants were asked to estimate the temperature of the room in which they sat. In both cases, the aggregate of estimates provided by the individuals outperformed the majority of individual responses. Later work by Gordon (1924) and Stroop (1932) explicitly tested the theory that aggregated crowd responses could outperform the average individuals’ response, with both studies finding strong evidence of crowd wisdom using different estimation tasks.

In early 2005, interest into the potential of crowd wisdom grew sharply following the release of Surowiecki’s (2005) book, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, which had inspired a new wave of research into forecast aggregation. Surowiecki’s book has been widely accredited with disseminating into popular culture the notion of the ‘wisdom of crowds’, by distilling and sharing findings from the technical literature to a much wider audience. Prior to his work, the idea of crowd wisdom, which has been established for over a century, seemed to have been largely

confined to a niche section of the technical literature.

Surowiecki's work provided much insight compared to these earlier studies into of the mechanism behind crowd wisdom. Surowiecki (2005) outlines four conditions necessary for crowd wisdom:

1. *Diversity of Opinion.* Individuals need to have their own private information.
2. *Independence.* Individuals' views aren't determined by the views of others in the crowd.
3. *Decentralisation.* Some individuals must have access to local or specialised knowledge.
4. *Aggregation.* A mechanism such as a decision maker is necessary for aggregating individual responses into a collective response.

Although these conditions appear to be somewhat overlapping – and therefore not theoretically distinct, Surowiecki's attempt to provide an account for the mechanisms underlying crowd wisdom formed a valuable foundation for subsequent, more rigorous research into the mechanisms of crowd wisdom.

1.2 Weighted Aggregation Approaches

Substantial research effort has been devoted to the development of better weighted aggregation approaches in recent decades. The basic idea of weighted aggregation approaches is to identify and leverage subsets of experts in the crowd (Budescu & Chen, 2015; Clemen, 1989; Cooke, 1991; Davis-Stober, Budescu, Dana, & Broomell, 2014; Mannes, Soll, & Larrick, 2014). In general, these models assume that differences in forecasters' expertise can be identified systematically through some characteristics of forecasters' responses, such as forecasters' confidences or the consistency of their responses. In particular, forecasters' past performance on a set of *seed questions* with known outcomes is often considered the most effective approach for deriving forecasters' weights (Budescu & Chen, 2015; Cooke, 1991). Early research on mathematical methods such as these is well-summarised in reviews by Clemen (1989); Clemen and Winkler (1999); Genest and McConway (1990), which taken together provide a comprehensive review of contributions from multiple fields

including psychology, forecasting, statistics, and social sciences, and discuss both theoretical and applied approaches to combining forecasts in the previous decades.

In Cooke’s (1991) seminal work, *Experts in Uncertainty: Opinion and Subjective Probability in Science (Environmental Ethics and Science Policy Series)*, he provides a detailed survey and critical exploration of the use of expert opinion in applied forecasting. Cooke includes many examples for combining expert opinion in a diverse range of real-world problems, including problems from artificial intelligence, risk analysis, military intelligence, aerospace engineering, and policy analysis. While the book covers a wide range of issues, such as how subjective probabilities should be collected and used in policy, Cooke also discusses various mathematical aggregation approaches. For example, he develops a classical model drawing from statistical hypothesis testing, where forecasters’ weights are derived from their calibration performance on a set of seed questions with known outcomes. Forecasters whose performance on the seed questions are below a theoretical threshold are assigned weights of zero, and their predictions are removed from the crowd. Cooke and Goossens (e.g., 2008) conduct a comprehensive review of the performance of the classical model in aggregating judgments from 45 expert panels across a range of domains, providing evidence of its efficacy.

More recent work by Mannes et al. (2014) examines the advantages of ranking and selecting expert forecasters based on their past performance. Like Cooke (1991), Mannes et al. (2014) demonstrated the advantages of aggregating over a select crowd of forecasters, rather than the whole crowd or the top-performing individual. In contrast to the older, more sophisticated weighted aggregation approaches in the preceding decades (Genest & McConway, 1990), Mannes et al. (2014) employed a simpler, equal-weighting approach to aggregating the responses of the selected crowd. They found that select crowds of the top-5 performing forecasters consistently outperformed the average forecast of the whole crowd in a wide range of possible settings over 90 archival datasets.

One challenge with using forecasters’ past performance as a measure of expertise arises when forecasters provide responses to different sets of seed questions (Atanasov et al., 2016; Budescu & Chen, 2015; Merkle, Steyvers, Mellers, & Tetlock, 2016). Seed questions often vary across multiple domains and difficulties, as such, differences in forecasters’ performance may simply reflect the fact that forecasters answered different sets of seed questions rather than reflecting true

differences in forecasters’ expertise. Past performance may therefore be an unreliable measure of expertise under these conditions.

One solution that has been proposed is to weight forecasters by their standardised scores, for example using a z-transformation (Atanasov et al., 2016; Merkle et al., 2016). Standardisation addresses this issue by rescaling forecasters’ weights based on how much they outperform other forecasters on the same question. Unfortunately, while standardised scores may be a better measure of expertise than raw scores, standardised models assume that differences in standardised units are commensurate across questions, which may not always be the case. For example, if expert and novice forecasters answered two different sets of questions, then a standardised unit improvement between experts would be more valuable than a standardised unit improvement in novices. Thus, although standardisation partly addresses the issue with estimating forecasters’ expertise on different sets of problems, it does not provide a complete solution.

The *Contribution-Weighted Model* developed by Budescu and Chen (2015) better addresses this issue by weighting forecasters by how much they improve the mean crowd prediction on seed questions. Budescu and Chen (2015) compared the performance of contribution-weighting to other weighting methods on two datasets comprising questions across multiple domains, including military, politics, business, policy, and economics, and found that forecasters’ contributions were much better measures of expertise than their absolute past performance. The contribution metric is effective because it rewards forecasters who are able to recognise when the crowd majority is incorrect and predict against the crowd, thus providing both expertise and diversity to the crowd since their predictions were generally anti-correlated with other forecasters’ predictions in the crowd. The contribution-weighting mechanism shares some similarity to work by Davis-Stober et al. (2014), who provide a general definition of the conditions necessary for crowd wisdom and formalises the trade-off between diversity and expertise in maximising crowd wisdom. Both Budescu and Chen’s (2015) and Davis-Stober et al.’s (2014) work highlight the importance of diversity between different forecasters in a crowd in maximising the improvement of the crowd prediction over individuals’ predictions, in contrast to traditional models that assign weights solely according to forecasters’ absolute past performance.

A common property in weighted aggregation approaches such as those proposed by Cooke (1991), Mannes et al. (2014), and Budescu and Chen (2015) is the removal of subsets of forecasters’

predictions from the aggregated prediction by assigning weights of zero to those forecasters. In assigning forecasters weights of zero, these models are assuming that those individuals will not improve the aggregate prediction, which may be potentially erroneous as these measures of expertise do not necessarily reflect forecasters' expertise in future predictions. Under conditions where forecasters expertise cannot be identified systematically, removing forecasters may simply reduce the crowd diversity, worsening the aggregated forecast. Partly for these reasons, aggregation approaches that do not rely on forecasters' past performance have become a major area of interest in recent years.

1.3 Cognitive Modeling Approaches

Cognitive modeling approaches have been proposed for identifying latent experts in the crowd and improving aggregated forecasts in similar decision problems (Lee & Danileiko, 2014; Lee, Steyvers, De Young, & Miller, 2012; Lee, Steyvers, & Miller, 2014). Lee and Danileiko (2014) showed that forecasters' probabilistic predictions could be aggregated using a hierarchical Bayesian cognitive model that jointly estimates the expertise and calibration of individuals' forecasts and the outcome of each event. In their model, forecasters' predictions are drawn from a Gaussian distribution centered on the true probability and the expertise of each forecaster is captured by the variance in the samples they draw from the Gaussian distribution. Expert forecasters can therefore be identified via greater precision in their knowledge of the true probability and thus lower variance in their draws. In their model, forecasters are also expected to differ in their level of calibration, with more calibrated forecasters reporting probability forecasts that better match the true probability, whereas poorly calibrated forecasters report probabilities that are systematically under-confident relative to the true probability.¹ An advantage of Lee & Danileiko's (2014) cognitive modeling approach is that it does not require the true question outcomes for any of the training problems to be known to the decision maker. The cognitive modeling approach is therefore appealing in problems where decision makers are unable to infer expertise based on forecasters' past performance or other characteristics of their responses. Nonetheless, the cognitive

¹The parameter in the calibration function estimated by the model take a linear-in-log-odds functional form with a single parameter capturing the magnitude of over- and under-estimation. This calibration function has a similar functional form to the recalibration function used in Chapters 3 and 4 of this thesis.

modeling approach may perform poorly when the majority of forecasters are biased. Since no ground truth information is provided to the model, it is unable to distinguish between questions where the majority of forecasters are incorrect vs. correct. Furthermore, as few comparisons between the cognitive modeling approach and these other expertise-based aggregation approaches have been made to date, it remains an open question whether the model will be effective in applied problems where it is hard to identify when the majority of individuals are biased.

1.4 Single-Question Aggregation Approaches

In early 2017, a paper published in *Nature* titled “A solution to the single-question crowd wisdom problem” proposed a novel aggregation algorithm that required no records of forecasters’ past performance (Prelec et al., 2017). Prior to this, decision makers who did not have records of forecasters’ past performance had relied on forecasters’ confidence as an indicator of expertise or combined crowd forecasts using simple aggregation rules such as the mean or median (Cooke, 1991). Prelec et al. showed that their Surprisingly Popular (SP) algorithm, which used forecasters’ meta-predictions about the proportion of other forecasters voting for each outcome, theoretically always identified the correct answer even when the crowd majority is incorrect. Prelec et al.’s work was revolutionary in that decision makers elicited forecasters’ meta-predictions as a means to improve the aggregated crowd forecast – an approach which had not been seen previously in the forecasting literature. Prelec et al.’s results thus showed substantial promise for improving forecasts in decision problems where decision makers often have no records of forecasters’ past performance. For example, Prelec et al. found that the SP algorithm was more effective than majority voting over a range of decision problems including judgments about the capital city of US states, general knowledge trivia questions, classification of malignant vs. benign skin lesions by dermatologists, and judgments about the prices of artworks by laypeople and professionals.

In a subsequent working paper (J. McCoy & Prelec, 2017) and in his PhD dissertation (J. P. McCoy, 2018), John McCoy extended the SP approach into multi-question contexts and probabilistic forecasting problems using a Bayesian and hierarchical approach, where the correct outcome is inferred as a latent variable from forecasters’ empirical responses using Bayesian inference. One drawback of this sophisticated modeling approach was that it required

stronger assumptions about the structure of information received by individuals and the ability of individuals in the crowd to make complex computations.

In the subsequent years, Prelec et al.’s (2017) paper had ushered in substantial research interest into forecast aggregation approaches that do not require past performance. In a study soon after, Lee et al. (2018) tested the efficacy of the SP algorithm on its ability to predict outcomes of games in the 2017-2018 American National Football League (NFL). In contrast to Prelec et al.’s findings on the efficacy of the SP algorithm, Lee et al. found that the SP algorithm offered no better forecasts than other simpler aggregation approaches. Despite this, the SP algorithm had shown reasonable calibration properties relating to the confidence of predictions to accuracy (Lee et al., 2018). Unfortunately, few other studies to date have investigated the mechanisms underlying the SP algorithm. Given the mixed results on the SP algorithm’s performance, it remains unclear the extent and conditions under which the SP algorithm can reliably outperform other aggregation approaches, such as majority voting or confidence-weighting.

Using a different paradigm, Palley and Soll (2019) developed an information-based aggregation approach that corrects for the bias in aggregated forecasts due to shared information between forecasters. Their *minimal pivoting* method uses forecasters predictions and meta-predictions about the average forecast of others to identify and correct for this bias. More generally, they developed a set of pivoting models that provide the optimal correction for the shared-information bias in the case of different idealised information structures in the crowd. Empirically, the minimal pivoting model was found to outperform simple averaging on both a grocery price estimation task and predictions for the outcome of NCAA basketball tournament games (although the improvement was not significant in the latter case). As the minimal pivoting approach has not been compared to other probabilistic forecasting approaches in the literature, it remains unknown whether the minimal pivoting approach can be expected to generally outperform these other approaches.

Most recently, Kurvers et al. (2019) developed a novel metric for identifying high-performing individuals in the crowd by using the similarity of their decision to that of other forecasters. The authors showed that in theory, when forecasters in the crowd are more often correct than incorrect, individuals with higher decision similarity to the crowd are always more accurate than individuals with lower decision similarity. Empirically, they found results consistent with their theoretical

model over a range of decision problems, including the classification of skin cancers and breast cancers, general knowledge trivia questions, and geopolitical forecasting problems. However, the authors also found that when the crowd majority is incorrect, the relationship between decision similarity accuracy was reversed, such that individuals with high decision similarity were also the least accurate. As such, decision similarity seems to be an unreliable metric on decision problems where the decision cannot predict in advance whether the crowd is more likely to be correct or incorrect. It therefore remains unclear the extent to which the decision similarity metric can be useful in practice, as decision problems often come with a high degree of uncertainty in its outcome and the decision maker cannot predict whether the crowd majority is going to be correct or incorrect. Furthermore, a formal algorithm for aggregating forecasts according to forecasters’ decision similarity remains to be developed. It’s therefore unknown how decision makers should aggregate forecasts using decision similarity and whether aggregation approaches based on forecasters’ decision are likely to be more effective than existing aggregation approaches in the literature.

1.5 Outline of the Present Research

Despite Prelec et al.’s (2017) remarkable findings on the Surprisingly Popular (SP) algorithm, very little is known about the environments in which the SP algorithm would be expected to perform well, relative to other aggregation approaches. The relationship between the SP algorithm and existing aggregation approaches in the literature has yet to be explored. The primary goal of Chapter 2 in this thesis is therefore to address these theoretical questions by identifying the mechanism by which the SP algorithm operates and to contextualise our understanding of that mechanism within the broader forecasting literature. In our results, we show that the SP algorithm in fact assigns weights to forecasters according to the absolute difference between forecasters’ votes and meta-prediction about the votes of other forecasters, and the SP algorithm therefore shares a similar functional structure to existing weighted-aggregation approaches in the literature.

While the SP algorithm’s weighting approach appear to be effective on binary decision problems, no study to date has tested the efficacy of the SP algorithm’s weights in probabilistic forecasting

problems. In Chapter 3, we therefore explore how knowledge of the SP algorithm’s mechanism can be used to improve probabilistic forecasting. First, we show that these weights are also effective for generating probabilistic forecasts, outperforming simple averaging, consistent with our theoretical predictions. Based on our theoretical model, we then develop a novel aggregation algorithm called the meta-probability weighting (MPW) algorithm, which derives weights in a way similar to the SP algorithm by using forecasters’ meta-predictions. We provide a detailed empirical comparison between the MPW algorithm and different probabilistic aggregation approaches in the literature across a range of forecasting problems and we show that the meta-probability weighting approach generally outperforms existing aggregation approaches in the literature.

While there has been considerable research effort into the development of aggregation approaches that do not require forecasters’ past performance, it remains unknown how these aggregation approaches are likely to perform relative to aggregation approaches outside the single-question context. In practice, decision makers may have the choice of improving forecasts by eliciting forecasters’ responses to questions with known outcomes (but on a domain which maybe unrelated to the domain of interest) and thus generating an estimate of forecasters’ expertise. In Chapter 4, we therefore examine the efficacy of the meta-probability weighting algorithm relative to algorithms that weight forecasters based on past performance on questions from the same question domain (within-domain weighting), or from a different question domain (cross-domain weighting). We demonstrate that cross-domain weighting is generally as effective as within-domain weighting and consistently outperforms aggregation approaches that do not use forecasters’ past performance, including the cognitive modeling approach developed by Lee and Danileiko (2014).

In each of these chapters, we provide a more detailed review of the relevant literature. We provide a brief summary and some concluding remarks in Chapter 5.

2 | Leveraging expertise using crowd meta-predictions

The work in this chapter was partly adapted for the manuscript Wilkening et al. (2020), which was written after this chapter. We have included a copy of this manuscript for reference in Section 6.1 of this thesis. For the technical details on the theoretical model referred to throughout Chapters 2 and 3 of this thesis, please see the attached manuscript.

The overarching aim of this chapter is to examine existing single-question aggregation approaches in the literature and identify how meta-predictions can be used to leverage expertise on binary decision problems. This chapter is structured as follows: first, we summarise the relevant literature on forecast aggregation, expertise-identification approaches in forecasting, and single-question aggregation approaches. We provide a formal definition of the existing single-question aggregation approaches in the literature and we propose a reformulation of the Surprisingly Popular (SP) algorithm (Prelec et al., 2017) that shows that it in fact weights forecasters by a measure of their latent expertise, which suggests that there is potential mechanism for the SP algorithm that has not been previously identified. We introduce a theoretical framework, adapted from Wilkening et al. (2020), that accounts for these results and generates predictions regarding the relationship between expertise and the performance of the SP algorithm relative to other aggregation approaches. Experiment 1 replicates Prelec et al.’s (2017) Study 1, and we conduct post-hoc analyses that demonstrate heterogeneity in forecasters’ meta-predictions. We examine the distributions of forecasters’ accuracy in each dataset and show that the SP algorithm performs significantly better in datasets where the distribution of forecasters’ accuracy is bi-modal and

forecasters' expertise is therefore heterogeneous. We conduct simulations where we manipulate the level of expertise, and show that the performance of the SP algorithm is strongly related to the overall level of crowd expertise. Experiment 2 tests for a non-monotonic relationship between expertise and the performance of the SP algorithm, and shows that the SP algorithm indeed offers the greatest improvement over other aggregation approaches on questions of moderate difficulty. Post-hoc analyses of Experiment 2 highlights key similarities between results for the current datasets and those in Experiment 1 and provides further evidence for the proposed SP mechanism. We reanalyse Lee et al.'s (2018) results in light of our findings, and we find some weak evidence in favour of the proposed SP mechanism in this dataset. We conclude with a discussion on the key contributions of the present research, in particular, our role in elucidating a fundamental mechanism of the SP algorithm that leverages latent expertise and connecting the current research to the wider forecasting literature on expertise-based aggregation approaches. We highlight potential extensions of the SP algorithm for probabilistic forecasting, which we explore in depth in the following chapter of this thesis.

2.1 Introduction

Human decision making often requires predicting the outcomes of future events and using the information that is available to find the best solution for complex problems. Forecast-aggregation models are commonly used to combine information from a crowd of individuals in order to improve upon the predictions from the individuals themselves. The phenomenon of improving crowd predictions through aggregating across individual predictions was termed the *Wisdom of Crowds* by Galton (1907), who provided one of the first rigorous demonstrations of crowd wisdom.

The Wisdom of Crowds has since been robustly demonstrated across a variety of domains, including meteorological forecasting (Murphy & Winkler, 1977), medical diagnoses (Meyer, Longhurst, & Singh, 2016), sports betting (Herzog & Hertwig, 2011; Simmons, Nelson, Galak, & Frederick, 2011), box-office success (Gillen, McKenzie, & Plott, 2018), predicting the outcomes of replications in scientific research (Dreber et al., 2015), and even subjective evaluations of music and film (Müller-Trede, Choshen-Hillel, Barneron, & Yaniv, 2017). In recent years, the Wisdom of Crowds has gained attention in political and economic forecasting where there are

often high stakes involved (Budescu & Chen, 2015; Mellers et al., 2015; Tetlock, 2017). Prediction markets are common platform that tries to harness the power of crowd wisdom but also incentivise forecasters to generate predictions by rewarding them based on their performance (Atanasov et al., 2016; Baillon, Tereick, & Wang, 2019; Wolfers & Zitzewitz, 2004). Nonetheless, aggregating decisions on problems may amplify deviations from rational models (Baillon, Bleichrodt, Liu, & Wakker, 2016), and therefore aggregation approaches may not be optimal in some problems.

2.1.1 Weighting by past performance

In recent decades, forecast-aggregation models have been developed to improve crowd predictions by identifying and leveraging the predictions of high-performing individuals or “experts” in the crowd over the predictions of low-performing individuals or “novices” (Budescu & Chen, 2015; Cooke, 1991; Davis-Stober et al., 2014; Mannes et al., 2014). In general, these models assume that differences in forecasters’ expertise can be identified systematically through some characteristics of forecasters’ responses. Forecasters’ past performance is often considered the best predictor of expertise, and forecasters’ performance on seed questions with known outcomes are commonly used to weight forecasters’ responses on questions of interest (Budescu & Chen, 2015; Cooke, 1991; Mannes et al., 2014).

In some domains, performance on seed questions may not effectively predict performance on the questions of interest, and therefore weighted aggregation models may offer no advantage over simple averaging (Genre, Kenny, Meyler, & Timmermann, 2013; Mannes, Larrick, & Soll, 2012). For example, Genre et al. (2013) examined expert forecasters’ predictions of GDP growth, inflation, and unemployment rate from the European Central Bank’s Survey of Professional Forecasters and found that aggregation methods based on past performance offer no advantage over the simple crowd average. Other studies have highlighted the fact that the top-performing traders in the stock market each quarter do not consistently outperform the average trader when performance is evaluated over several periods of time (Mannes et al., 2012). In these environments, performance may be determined mostly by luck or other random factors, such that forecasters’ past performance may not accurately reflect expertise, and therefore expertise-based models tend to offer little benefit over simpler aggregation approaches such as the majority voting.

In practice, decision makers often do not have access to records of forecasters' past performance on a set of relevant seed questions, for example, because the time and costs required for obtaining this information. As aggregation approaches that rely on forecasters past performance are not available in these environments, researchers have since developed aggregation approaches that can generate accurate forecasts without using this information.

2.1.2 Confidence is an unreliable predictor of expertise

Researchers have sought to identify forecasters' expertise by other characteristics of their responses, such by forecasters' level of confidence, when records of forecasters' past performance is unavailable. Confidence-based approaches treat confidence as a predictor of expertise, assigning greater weights to more-confident judgments than less-confident judgments in the aggregation process. For example, Cooke (1991) examines various linear and non-linear confidence-weighting approaches. One problem with weighting forecasters by confidence is that the relationship between confidence and expertise depends strongly on task difficulty (Hertwig, 2012). Confidence has been shown to be positively correlated with accuracy when task difficulty is low and negatively correlated with accuracy when task difficulty is high (Koriat, 2008). As a result, confidence-weighting methods will tend to perform poorly only on difficult tasks where other standard aggregation approaches such as majority voting would also tend to fail. The literature is rife with examples where confidence is a poor predictor of forecaster accuracy (Fischhoff & MacGregor, 1982; Koriat, 2008, 2012; Prelec et al., 2017).

Findings from social psychology also support the notion that confidence is generally a poor measure of expertise. The correlation between self-assessments of skill and actual measures of performance are typically moderate at best (Dunning, Heath, & Suls, 2004; Kruger & Dunning, 1999). Similarly, assessments of others' expertise often do not correspond to their true skill (Lovallo & Kahneman, 2003; Yaniv & Kleinberger, 2000). For example, when forecasters are given knowledge about others' predictions on a task requiring them to estimate the historical dates of events, forecasters almost always place greater weight in their own predictions than others', even in cases where others' predictions are shown to be superior to their own (Yaniv & Kleinberger, 2000). Similarly, Lovallo and Kahneman (2003) describe a survey of 1 million

students conducted by the US College Board in the 1970s that showed that a large majority of high school students believe they are above-average in skill across social and achievement domains, including leadership, athletic prowess, and social ability, with less than 10% of respondents considering themselves below-average on a number of these domains (Dunning, Meyerowitz, & Holzberg, 1989; Lovall & Kahneman, 2003). More generally, people’s tendency to see themselves more favourably than in reality, often called a self-enhancing bias, has been robustly demonstrated in the psychology literature (Krueger, 1998; Kruger & Dunning, 1999). For forecasting problems, such discrepancies between forecasters’ self-perception and reality can result in poor performance for confidence-based aggregation approaches due to the fact these algorithms rely on confidence as a measure of expertise.

2.1.3 Differences in meta-cognitive knowledge between experts and novices

Findings from social psychology suggest that forecasters’ meta-cognitive knowledge may be a useful measure of expertise. Experts and novices are shown to differ systematically in their awareness of their own level of skill and the level of skill of others (i.e., meta-cognitive knowledge). In their classic study, Kruger and Dunning (1999) demonstrated that low-skilled individuals tended to exhibit the greatest over-confidence in their own ability in tests of humour, grammar, and logic, while experts tended to have an accurate perception of their own abilities. Their findings have since been well-replicated in a range of other domains (Dunning et al., 2004, 1989).

Similar findings have also been made in other fields. For example, Eteläpelto (1993) compared expert and novice computer programmers in terms of their meta-cognitive awareness and found that expert programmers demonstrated much better meta-cognitive knowledge of parts of the task they were struggling with, compared to novice programmers. Expert programmers often provided specific, detailed reasons why they did not understand certain parts of the programming task, whereas novices provided responses at a more general, diffuse level (Eteläpelto, 1993). Similarly, Bromme, Rambow, and Nückles (2001) found that laypersons demonstrated much higher bias than computer experts in their estimates of the distribution of Internet concepts and general knowledge concepts among students. Experts were thus aware that they themselves possessed

specialised knowledge not available to non-experts, whereas laypersons were generally not privy to such meta-cognitive knowledge.

Altogether, these findings suggest that the skills that engender competence are often the same skills necessary to be able to judge self-competence (Kruger & Dunning, 1999). Forecasters’ meta-cognitive knowledge may therefore be useful as a measure of expertise in environments where forecasters’ past performance is unavailable.

2.1.4 The Surprisingly Popular algorithm

Prelec et al. (2017) proposed the Surprisingly Popular (SP) algorithm for aggregating crowd forecasts in single-question forecasting problems. The SP algorithm generates predictions using forecasters’ predictions and *meta-predictions* – an estimate of the percentage of other people that would endorse a particular response. The SP algorithm has the remarkable ability to identify the correct answer even when the majority of forecasters are incorrect. The algorithm achieves this by selecting the answer that is more popular than predicted by comparing the average crowd prediction with the predicted crowd endorsement (i.e., the mean meta-prediction of the crowd).

Adopting the example from Prelec et al. (2017), suppose that you have no knowledge of US geography and you have responses from a crowd of forecasters for the statement “Philadelphia is the capital of Pennsylvania”. Forecasters have been asked (1) whether this statement is more likely true or false, (2) for the probability (i.e., confidence) that this statement is true, and (3) for their meta-prediction of what percentage of other people will believe the statement is “true”. How then should you aggregate these responses to identify the correct outcome? In practice, about two-thirds of forecasters will believe the statement to be true and are therefore incorrect (since Harrisburg is the capital city of Pennsylvania; Prelec et al., 2017). Confidence-based algorithms also tend to predict incorrectly as both correct and incorrect forecasters are equally confident that they are correct (Prelec et al., 2017). In comparison, the crowd’s average meta-prediction is found to consistently overestimate the proportion of true voters, and the SP algorithm is therefore able to produce the correct prediction by utilising forecasters’ meta-knowledge about what other forecasters will predict.

The theoretical model proposed by Prelec et al. (2017) provides an intuitive explanation for

how meta-predictions can be used to generate correct predictions. The SP algorithm is formalised as a Bayesian model where forecasters have a shared prior belief about the outcome of each question, and they update their beliefs according to the private signals they receive regarding the likelihood of different outcomes in factual and counter-factual worlds. Forecasters receive signals which can be conceptualised as identical and independent tosses from a biased coin. For binary questions with two outcomes (i.e., the cases considered exclusively in this thesis), there are two possible biased coins – one for the correct outcome and one for the incorrect outcome. Forecasters are assumed to know the biases of these coins, but not which ones correspond to the correct or incorrect outcomes. Knowledge of these coin biases allows forecasters to compute the proportion of forecasters who would predict a given outcome in each of the factual and counter-factual worlds. The average meta-prediction of forecasters in the real world, which lies somewhere between these hypothetical factual and counter-factual worlds, would therefore also always underestimate the correct outcome, leading the SP algorithm to identify the correct outcome.

Here is an intuitive example of how the model operates. For the statement “Philadelphia is the capital of Pennsylvania”, a given crowd of forecasters might believe that 70% of forecasters would predict the statement to be true in the factual world (where Harrisburg is the capital city, and the statement is therefore false). In the counter-factual world, where the statement is actually true, they might believe that 90% of forecasters would predict the statement to be true. Unfortunately, forecasters do not know which world is the factual world and therefore do not know whether the statement is true or false. In order to generate a meta-prediction in the real world, forecasters need to combine their beliefs about the proportion of votes in both factual and counter factual worlds based on their beliefs about the probability of each world being factual. As a result, the exact value of their meta-prediction would therefore lie strictly between 70% and 90%. Since the statement is factually false, the proportion of forecasters voting true would converge to 70%, but the average crowd meta-prediction would always be equal or strictly greater than 70%. As the mean meta-prediction is greater than the proportion of true votes, the SP algorithm would predict the answer to be false, thus predicting correctly even for questions where a majority of 70% of forecasters in the crowd predict incorrectly and contrary to the SP prediction.

Empirically, Prelec et al. (2017) found that the SP algorithm generally outperformed majority voting but not confidence-based aggregation algorithms across a variety single-question forecasting

problems, including general-knowledge questions about geography, science, history, and language; classifications of skin lesions by professional dermatologists; and estimates of art prices by novices and art experts. Their results demonstrated that the SP algorithm was generally effective in most domains, but particularly for US states questions about the capital city of each state, where it was found to provide up to a 30% increase in accuracy over majority voting, and almost 15% over confidence-based models. Both theoretically and empirically, Prelec et al.'s (2017) findings thus constituted an impressive development in the single-question forecasting literature.

More recently, Lee et al. (2018) examined the ability of the SP algorithm to predict the outcomes of games in the 2017-2018 US National Football League (NFL) season. Comparing 256 NFL games over a period of 17 weeks, Lee et al. (2018) found that the SP algorithm offered no significant improvement over other aggregation approaches including majority voting and confidence-weighting. Lee et al. (2018) is the second published study in the literature to examine the SP algorithm, and the first study to examine its performance on questions of genuine predictions (i.e., about the outcomes of future events). While Lee et al. (2018) found that the SP algorithm did not outperform any other aggregation approaches, they demonstrated that the SP algorithm has reasonable calibration properties in terms of confidence vs. accuracy. Specifically, they found that the accuracy of the SP algorithm increased as the difference between the average crowd prediction and the mean meta-prediction increased, and therefore could potentially be used to derive an estimate of uncertainty for events. Such a measure of uncertainty would be useful in extending the SP algorithm in the domain of probabilistic forecasting where the SP algorithm is currently not applicable. Nonetheless, as Lee et al. (2018) noted, the exact scaling and transformation used to obtain such estimates remains to be developed theoretically. Chapter 3 examines another potential way of extending the results from the current chapter to generate probabilistic forecasts, by using the normalised absolute difference between forecasters' predictions and meta-predictions as linear weights, which may be useful in forecasting problems where continuous predictions are preferred to discrete binary predictions.

2.1.5 Questions to be addressed

Despite the SP algorithm’s impressive theoretical properties, its empirical performance relative to other aggregation approaches seems to vary substantially from one domain to another. For example, on datasets regarding the capitals of US states, the SP algorithm offers up to 15% increase in accuracy over the next-best algorithm, whereas on other datasets, such as for predicting the outcomes of NFL games, it is slightly less accurate than majority voting. The theoretical model developed by Prelec et al. (2017) currently offers no mechanism to explain such differences in results across domains. Specifically, what are the kinds of environments that we expect the SP algorithm to outperform other algorithms, and what defines the environments where it offers no advantage over other algorithms? This knowledge would be useful for practitioners seeking to apply the algorithm to novel forecasting problems, due to the fact that the SP algorithm requires eliciting forecasters’ meta-predictions, and thus require potential additional costs for elicitation. Naturally, these practitioners would prefer not to expend such costs in domains where majority voting would suffice.

Secondly, the relationship between the SP algorithm and other aggregation approaches in the forecasting literature remains to be explored. As discussed, many existing forecast-aggregation approaches in the literature operate by identifying and extracting expertise in the crowd, whether by using forecasters’ confidence, past performance, or some other characteristic as a measure of expertise (Budescu & Chen, 2015). An important question is whether the SP algorithm could be operating under a similar expertise-driven mechanism. As we know, meta-cognitive knowledge is an indicator of expertise (Dunning et al., 1989), and thus, it is possible that the SP algorithm could be using forecasters’ meta-predictions to leverage latent expertise in a similar fashion as these other algorithms. Neither Prelec et al. (2017) nor Lee et al. (2018) have explored the relationship between the SP algorithm and these existing expertise-based approaches, thus the relationship between these approaches remains unclear.

In the present chapter, we report results from two experiments that explore how the performance of the SP algorithm relates to expertise. In our results, we show that the performance of the SP algorithm, relative to the other algorithms, is related non-monotonically to the proportion of experts in the crowd. When the crowd comprises a moderate share of experts, the algorithm

performs well relative to the alternative algorithms. However, when the share of experts is large or small, the SP algorithm performs poorly relative to the alternatives.

2.2 Preliminaries

2.2.1 Individual responses and predictions

In this section, we provide a formal definition of the notation used throughout this chapter.

We consider a series of K events indexed by $k = \{1, \dots, K\}$ each with a binary outcome $o_k \in \{T, F\}$. We say an event is true if $o_k = T$. For each event k , a crowd of N_k forecasters is assembled. From each forecaster $i \in \{1, \dots, N_k\}$ we elicit three reports: the forecaster's prediction of whether an event is true, $V_{i,k} \in \{0, 1\}$, the forecaster's estimate of the probability that the event is true $P_{i,k} \in [0, 1]$, and the forecaster's meta-prediction, which is the forecaster's prediction of the proportion of the crowd who predict that the event is true, $M_{i,k}^V \in [0, 1]$. We let $X_{i,k} := (V_{i,k}, P_{i,k}, M_{i,k}^V)$ be forecasters i 's full report and assume that the reports are consistent in the sense that $V_{i,k} = 0$ if $P_{i,k} < 0.5$ and $V_{i,k} = 1$ if $P_{i,k} > 0.5$. The way we elicited the predictions ensured that this assumption was never violated. We assume that if $P_{i,k} = 0.5$, forecasters vote randomly for either outcome with equal probability.

We also assume that forecasters' probabilistic forecasts are a direct combination of their votes and their confidences, such that as their subjective confidence increases, forecasters probabilistic predictions would tend towards 1 or 0. We elicited forecasts in this indirect manner, rather than asking forecasters to provide confidences on an arbitrary scale so that their responses could be evaluated directly using proper scoring metrics, which we discuss in more detail in the following chapter.

2.2.2 Measures of binary classification performance

Let $X_k = \{X_{i,k}\}_{i=1}^{N_k}$ be the full set of reports for event k . Each algorithm we consider is a mapping $T : X_k \rightarrow \{0, 1\}$, which aggregates the data from a single event into a categorical forecast of whether the event is true or false. As discussed above, we restrict attention to single-question algorithms which use only the reports for event k to form the forecast for that event.

In this chapter, we compare the performance of algorithms on using both Cohen’s Kappa Coefficient and percentage accuracy. Although accuracy is the simplest and most intuitive measure of an algorithm’s (or an individual’s) ability to predict the correct outcome, it provides a poor measure of algorithm performance when classification tasks are imbalanced (i.e., when there are unequal proportions of “true” vs. “false” outcomes). Cohen’s Kappa Coefficient is considered a more robust measure than accuracy as it takes into account the possibility of the agreement occurring by chance (Cohen, 1960). We thus use Cohen’s Kappa Coefficient as the main measure of performance, but also report percentage accuracy for completeness, and due to its ease of interpretation.

Throughout this chapter, we assess statistical significance between predictions of different algorithms using classical Frequentist 95% confidence intervals (CIs), which indicate a significance difference when the null hypothesis value (for all comparisons in this chapter, $H_0 = 0$) is not contained within the interval. In order to compare predictions between algorithms on the same set of events, we exclude any events *pairwise* for a particular algorithm when that algorithm makes a non-prediction (see subsection 2.2.4 below, where we define each algorithm). We then compute 95% confidence intervals for paired differences in Cohen’s Kappa between the SP algorithm and each other algorithm using the bias-corrected and accelerated bootstrap (Efron, 1987). The bias-corrected and accelerated bootstrap is non-parametric, and is considered generally more robust than equivalent parametric approaches such as the paired-samples *t*-test (Efron, 1987).

To compute each interval, bootstrap samples of size n are drawn from the original sample of n events with replacement, such that each bootstrap sample may contain a distribution of events that differ from the original sample. The classification performance of each algorithm is then assessed on each bootstrap sample, and a distribution of performance statistics are constructed, after an acceleration and bias-correction factor is applied (see Efron, 1987, for details). The 95% confidence-interval limits are then obtained by computing the 2.5% and 97.5% percentiles of the resulting distribution.

2.2.3 The Surprisingly Popular algorithm

Our main algorithm of interest is the Surprisingly Popular (SP) algorithm of Prelec et al. (2017), which generates a binary categorical prediction for each event. We first show that the SP algorithm can be rearranged such that each forecaster's vote is weighted by the normalised, absolute difference between their vote and meta-prediction. We begin with the original form of the SP algorithm and rearrange it to show that it is identical to a reformulation that we will examine throughout this chapter, which takes a weighted form. The weights in our reformulation of the SP algorithm are given by the absolute difference between their vote and their meta-prediction, normalised by the sum of this difference over all forecasters:

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{|V_{i,k} - M_{i,k}^V| |V_{i,k}|}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (2.1)$$

In the original SP algorithm, the proportion of the crowd voting for that outcome is compared to the mean meta-prediction, and the most under-predicted outcome is then predicted to be correct. Formally,

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} (V_{i,k} - M_{i,k}^V) > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2.2)$$

The crowd for an event comprising N_k forecasters can be decomposed into T_k forecasters who vote true and F_k forecasters who vote false, $N_k = T_k + F_k$. The report of each forecaster who votes true for event k , $t_k \in \{0, \dots, T_k\}$, is given by $X_{t,k} := (V_{t,k}, P_{t,k}, M_{t,k}^V, M_{t,k}^P)$, and the report of each forecaster who votes false for event k , $f_k \in \{0, \dots, F_k\}$, is given by $X_{f,k} := (V_{f,k}, P_{f,k}, M_{f,k}^V, M_{f,k}^P)$. The SP equation can therefore be decomposed into

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} (V_{t,k} - M_{t,k}^V) + \sum_{f=1}^{F_k} (V_{f,k} - M_{f,k}^V) > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

Rearranging this, we get

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} (V_{t,k} - M_{t,k}^V) > - \sum_{f=1}^{F_k} (V_{f,k} - M_{f,k}^V) \\ 0 & \text{otherwise.} \end{cases} \quad (2.4)$$

As $V_{f,k} = 0$, $V_{t,k} = 1$, and $M_{i,k}^V \in [0, 1]$, the difference between votes and vote meta-predictions for any individual who votes false will always be equal to or less than 0,

$$V_{f,k} - M_{f,k}^V \leq 0, \quad (2.5)$$

and the difference between votes and vote meta-predictions for any individual who votes true will always equal or exceed 0,

$$V_{t,k} - M_{t,k}^V \geq 0. \quad (2.6)$$

The SP equation is therefore equivalent to

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} |V_{t,k} - M_{t,k}^V| > \sum_{f=1}^{F_k} |V_{f,k} - M_{f,k}^V| \\ 0 & \text{otherwise.} \end{cases} \quad (2.7)$$

Adding the terms on the left to both sides, we obtain

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} 2|V_{t,k} - M_{t,k}^V| > \sum_{t=1}^{T_k} |V_{t,k} - M_{t,k}^V| + \sum_{f=1}^{F_k} |V_{f,k} - M_{f,k}^V| \\ 0 & \text{otherwise.} \end{cases} \quad (2.8)$$

Since

$$\sum_{t=1}^{T_k} |V_{t,k} - M_{t,k}^V| + \sum_{f=1}^{F_k} |V_{f,k} - M_{f,k}^V| = \sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|, \quad (2.9)$$

we can collect the terms on the right of equation 2.8:

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} 2|V_{t,k} - M_{t,k}^V| > \sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V| \\ 0 & \text{otherwise.} \end{cases} \quad (2.10)$$

After dividing both sides by the RHS term and dividing both sides by 2, we obtain

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \frac{\sum_{t=1}^{T_k} |V_{t,k} - M_{t,k}^V|}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (2.11)$$

which is identical to

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} \frac{|V_{t,k} - M_{t,k}^V|}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (2.12)$$

Since $V_{t,k} = 1$, we can multiply both sides by $V_{t,k}$ and simplify the terms on the right to obtain

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} \frac{|V_{t,k} - M_{t,k}^V| V_{t,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (2.13)$$

As $V_{f,k} = 0$,

$$\sum_{f=1}^{F_k} \frac{|V_{f,k} - M_{f,k}^V| V_{f,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} = 0, \quad (2.14)$$

and we can add this summation term to both sides of the previous equation and simplify the terms on the right to obtain

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T_k} \frac{|V_{t,k} - M_{t,k}^V| V_{t,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} + \sum_{f=1}^{F_k} \frac{|V_{f,k} - M_{f,k}^V| V_{f,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (2.15)$$

Collecting the terms on the left, we obtain the weighted version of the SP algorithm, thus proving

that the two versions of the SP algorithm are equivalent,

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{|V_{i,k} - M_{i,k}^V| V_{i,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (2.16)$$

We note that both formulations produce identical predictions. Our rearrangement of the SP algorithm highlights how individuals in the crowd are weighted, whereas the original formulation is perhaps simpler to understand. The weights in our reformulation allow us to directly identify the difference in expertise between individuals in the crowd, whereas such analysis is not directly available in the SP algorithm’s original formulation. The utility of this reformulation is demonstrated in the analyses throughout this chapter.

2.2.4 Other categorical forecast-aggregation algorithms

We also compute categorical predictions for a number of alternative algorithms that are popular in the literature and were originally compared to the SP algorithm in Prelec et al. (2017). As Table 2.1 shows, our comparison set includes three other classification algorithms: the majority vote algorithm, the confidence-weighted algorithm, and the max-confidence algorithm.

The majority vote algorithm predicts that the answer with the most votes is going to be correct, without taking into account forecasters’ meta-predictions or uncertainty in forecasters’ predictions. The confidence-weighted and max-confidence algorithms both take into account uncertainty in forecasters’ predictions but in different ways: the confidence-weighted algorithm weights forecasters’ votes by their confidences, or alternatively, assigns equal weights to forecasters’ probabilistic predictions; in contrast, the max-confidence algorithm calculates the average confidence separately for those who vote “true” and those who vote “false”, effectively allowing one highly-confident individual who votes for a particular answer to dictate the aggregate prediction over a larger group of less-confident, dissenting individuals.

For readability, we have omitted from the Table the special case that both sides of the inequality for a particular algorithm are equal: for majority voting, if there were equal votes for both “true” and “false”; for confidence-weighting, if both “true” and “false” voters were equally confident; for the max-confidence algorithm, if the average confidence for both “true” and “false”

votes were equal; and for the SP algorithm, if the proportion of the crowd voting “true” is exactly equal to the mean meta-prediction. In these cases, these algorithms would generate no prediction and their performance would be calculated on the remaining problems in the dataset where they do generate predictions. Out of the total 940 problems across the 11 datasets analysed in Experiments 1 and 2, there were a total of 10 events with a non-prediction for the majority vote algorithm, 3 for the confidence-weighted algorithm, 7 for the max-confidence algorithm, and 0 for the SP algorithm. We analysed the results in this manner in order to avoid assigning random predictions of true or false (with equal probability) for each tie, which would unfairly boost the performance of some algorithms over others depending on randomisation. Nonetheless, this meant that each algorithm was evaluated on a very slightly different set of problems. Given the large set of problems analysed, we would not expect these exclusions to have any meaningful impact on our results or conclusions.

Table 2.1: Formula and Descriptions of Each Aggregation Algorithm. See Section 2.2.1 for Details on Notation.

Algorithm Name	Formula	Description
Majority Vote	$T_{MV}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{V_{i,k}}{N_k} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$	Select the answer with the most votes
Confidence-weighted	$T_{CW}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{P_{i,k}}{N_k} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$	Select the answer with the greatest confidence-weighted vote
Max-confidence	$T_{MC}(X_k) = \begin{cases} 1 & \text{if } \sum_{\{i V_{i,k}=1\}} \frac{P_{i,k}}{N_{t,k}} > \sum_{\{i V_{i,k}=0\}} \frac{1-P_{i,k}}{N_{f,k}} \\ 0 & \text{otherwise.} \end{cases}$	Select the answer with the greatest mean confidence (Note: $N_{t,k} = \sum_{i=1}^{N_k} V_{i,k}$; $N_{f,k} = N_k - N_{t,k}$)
Surprisingly Popular	$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{ V_{i,k} - M_{i,k}^V V_{i,k}}{\sum_{j=1}^{N_k} V_{j,k} - M_{j,k}^V } > 0.5 \\ 0 & \text{otherwise.} \end{cases}$	Select the answer that is surprisingly popular

2.3 A theoretical model for understanding the SP mechanism

In this section, we propose a new theoretical framework for understanding how forecasters' meta-predictions can be used to distinguish between latent experts and novices in the crowd, and thus can be used to leverage expertise. In Wilkening et al. (2020)¹, we develop a theoretical model in which forecasters draw signals from different information systems that are ordered in terms of informativeness. We find that the expected difference between the vote of an expert and their meta-prediction is larger than the expected difference between the vote of a novice and the novice's meta-prediction in a large class of decision problems. As can be seen in our formulation of the SP algorithm above, a forecaster's contribution to the decision made by the algorithm (i.e., the weight assigned by the SP algorithm to that forecaster's vote) is proportional to the difference between their vote and meta-prediction. Thus, our theoretical model predicts that the SP algorithm generally assigns greater weights to experts than novices for a large class of decision problems. Below, we provide basic outline of the theoretical model.

We consider a Bayesian model in which a crowd of N forecasters is assembled to predict the outcome of a single event. The outcome of the event, $o \in \{T, F\}$, is binary and can be true or false. Forecasters share a common prior $p(T)$ that the event is true, and the prior can either be *unbiased*, $p(T) = .5$, or *biased*, $p(T) \neq .5$.

Each forecaster receives a private signal s , that is a random variable taking on real value realisations in the set $\{s_1, \dots, s_m\} \cup \{s_\emptyset\}$ where $0 \leq s_1 < s_2 < \dots < s_m \leq 1$ and $s_1 < s_\emptyset < s_m$. As our outcome space is binary, it is without loss of generality that we normalise the signals so that their value is equal to the posterior belief that an event is true. That is, $s_j := p(T|s_j)$. We let $s_\emptyset$ represent the case where an individual receives an uninformative signal so that $s_\emptyset := p(T)$.

We use a left stochastic matrix called an *information service* to model the distribution of signals across forecasters in each state.² Participants receive signals from an information system denoted as Q . We assume that the properties of Q are common knowledge to all forecasters.

¹the manuscript is included in the Appendices – Section 6.1

²See Blackwell (1953); Blackwell and Girshick (1979); Marschak and Miyasawa (1968); Marschak and Radner (1972) for general treatments of information systems.

An information service is composed of a likelihood matrix $[Q_{oj}]_{2 \times (m+1)}$. Each element of the first row of Q represents the probability that the signal is s_j given the outcome is $o = T$. Likewise, each element of the second row of Q represents the probability that the signal is s_j given the outcome is $o = F$. For ease, we will denote the first row elements with T and the second row elements with F . Thus $Q_{Tj} := Q_{1j} = p(s_j|T)$ while $Q_{Fj} := Q_{2j} = p(s_j|F)$.

We note two important features of an information service. First, an information service acts as a transition matrix from a state of nature to a signal and thus $\sum_j Q_{oj} = 1$ for each row $o \in \{T, F\}$. Second, upon receiving a message from an information service, agents revise their priors using Bayes rule. For any signal that occurs with positive probability (i.e., where $Q_{Tj} + Q_{Fj} > 0$), the posterior belief that the event is true is given by

$$p(T|s_j) = \frac{p(T)Q_{Tj}}{p(T)Q_{Tj} + p(F)Q_{Fj}}.$$

By construction, this is equal to s_j for all signals that occur with positive probability.

This framework is useful because it allows us to model how differences in forecasters' latent expertise impact forecasters' meta-predictions about the forecasts of others in the crowd. Forecasters differ in their expertise based on the private signals they receive, with a more expert forecaster receiving a more informative private signal than a novice forecaster. Specifically, we define informativeness by:

Definition 1 *Forecaster i has a **more informative private signal** than forecaster j if either (i) $s_i < s_j < s_\emptyset$ or (ii) $s_i > s_j > s_\emptyset$.*

Intuitively, the informativeness of a forecasters private signal is related to the distance between his posterior and the common prior. We have restricted attention to cases where s_i and s_j are either both greater than $s_\emptyset$ or both less than $s_\emptyset$ so that distance is directly related to the relative changes in the likelihood ratios of the two forecasters.³

We note that the ordering of private signals is related to the extremity of the posterior (from $p = .5$), but is not equivalent to extremity in decision problems where there is a biased prior.

³For example, if $s_i > s_j > s_\emptyset$, then $\frac{Q_{Ti}}{Q_{Fi}} > \frac{Q_{Tj}}{Q_{Fj}} > \frac{p(T)}{p(F)}$. Thus $|s_i - s_\emptyset| > |s_j - s_\emptyset|$ implies $|\frac{Q_{Ti}}{Q_{Fi}} - \frac{p(T)}{p(F)}| > |\frac{Q_{Tj}}{Q_{Fj}} - \frac{p(T)}{p(F)}|$.

For example, in a problem where the common prior is $s_\emptyset = .75$, a forecaster who has a signal of $s_i = 0.5$ will have received a more informative signal than a forecaster with a signal of $s_j = 0.6$. This distinction is important because the absolute magnitude of a forecaster’s posterior beliefs that an event is true (i.e., the distance from .5) can be a poor indicator of expertise in biased problems. For example, in the example where the common prior is $s_\emptyset = .75$, if the true outcome probability is .5, then a forecaster who has a posterior of .5 has received more information, and is therefore a greater expert than a forecaster who has a posterior of .6, despite the latter having a more extreme posterior.

From the construction of this simple framework and the assumption that the properties of the information service is known to all forecasters, we can compute the prediction and meta-prediction for any individual forecaster, conditioned on the signal they receive. This property extends to the case where expert and novice forecasters may receive different information services, which we discuss in more detail in Wilkening et al. (2020). In order for a forecaster’s weight in the SP algorithm (i.e., the absolute difference between a forecaster’s vote and the forecaster’s meta-prediction about the votes of others) to be greater for experts than novices, three additional assumptions are required: (1) there is only a small proportion of experts in the crowd, (2) novices’ information service is a *strict garbling* of experts’ information service, and (3), forecasters’ information services are symmetric.⁴ Our theoretical framework provides a plausible mechanism by which the SP algorithm may be leveraging expertise.

Although our theoretical framework is not statistical in nature (i.e., we do not model the random variation in forecasters’ responses due to noise), it can provide an intuitive explanation for the different patterns of results between datasets observed below in Experiment 1. As the informativeness of experts’ and novices’ information systems become increasingly similar, the magnitude of differences between forecasters’ votes and meta-predictions (i.e., the relative contributions of experts vs. novices) become decreasingly small. It follows that random noise in forecasters’ responses would have an increasingly larger effect on the observed differences between forecasters’ votes and meta-predictions as experts’ and novices’ information systems become increasingly similar. These theoretical predictions generated from our model are able to account

⁴See Appendix B of Wilkening et al. (2020) – attached in Section 6.1 – for a formal definition of these terms, and proofs for these propositions.

for the pattern of results observed empirically. Conversely, our theoretical model predicts that we would observe the SP algorithm to be most effective when crowd expertise varies strongly, such that (1) crowds contain both experts and novices, and (2) experts receive much stronger signals than novices (i.e., rather than slightly stronger signals). As we show below in Experiments 1 and 2, these predictions are also entirely consistent with the results observed empirically.

2.4 Experiment 1

Experiment 1 replicated Prelec et al.’s (2017) study 1, which asked true or false questions about the capital cities of US States. We selected this study as it utilised the question set where the SP algorithm had the best performance in Prelec et al.’s (2017) original study, thus, it was a natural environment to study the mechanisms underlying the SP algorithm’s performance. We also used a larger sample size in order to analyse patterns of responses within particular subsets of the crowd.

2.4.1 Methods

We conducted the experiment online, with all participants recruited using Amazon Mechanical Turk. In Prelec et al.’s (2017) experiments, forecasters were monetarily incentivised for accurately predicting the outcome as well as accurately predicting the proportion of the crowd endorsing each response. As our experiment was performed online, we removed the financial incentives for accurate forecasts or meta-predictions to reduce the likelihood of participants looking up the answers. We tested 60 respondents recruited from Amazon Mechanical Turk and only respondents inside the US were able to participate. Each survey was administered online using Qualtrics, and participants were paid a flat fee of US \$2.00 for completing the survey. Participants were asked to answer each question as honestly as they could, and were asked not to cheat (e.g., by looking up any of the questions online). Eight individuals who reported cheating at the task or had failed to complete the survey were excluded from the analyses, but were still paid. We completed data collection in May 2017 and analyses were conducted on the data of the remaining 52 participants. The list of questions in our experiment are included in the Appendices section (see Section 6.1). See Prelec et al. (2017) for details on the datasets they collected.

The survey consisted of 50 trials (one for each US state, in alphabetical order of state). An example of a trial is shown in figure 2.1. On each trial, participants were shown the sentence “X is the capital of Y” where X was the most populous city in the state Y. For example, on the first trial, all participants saw the bolded statement “**Birmingham is the capital of Alabama.**” For each statement, participants were asked to answer three questions:

1. Is this statement more likely to be true or false?
2. What is your estimated probability of being correct? (50 to 100 percent)
3. What percentage of other people do you think thought the bolded statement was true? (0 to 100 percent)

Participants were also asked to answer two other questions about their beliefs under the hypothetical scenarios of the answer being true or false. Specifically, they were asked to provide estimates of the proportion of the crowd that will predict the outcome to be true in a hypothetical world where the answer was true, and a hypothetical world where the answer was false:

1. Imagine if the bolded statement was in fact true. What percentage of people would correctly think it was true? (0 to 100 percent)
2. Imagine if the bolded statement was in fact false. What percentage of people would incorrectly think it was true? (0 to 100 percent)

In a Bayesian framework, the law of total probability implies that an individual’s meta-prediction will be bounded below by their meta-prediction when the state is false and bounded above by their meta-prediction when the state is true. However, in our data, an individual’s meta-prediction fell outside the interval generated by the additional questions in 59.9 percent of cases. The meta-prediction matched both bounds in an additional 10.3 percent of cases. These results suggest that subjects may not understand how to answer these additional questions or find it difficult to make meta-predictions in counter-factual worlds.

We had intended to use the law of total probability to generate an additional measure of an individual’s belief that the statement is true. However, given that only 29.8 percent of our data would generate a consistent measure, we chose not to use the additional questions in our analyses.

We analysed the data from this experiment alongside data from the five datasets collected by Prelec et al. (2017) for which they collected confidences or probability forecasts (see Prelec et al., 2017, for details). We refer to our dataset as “States (MWH)” in order to distinguish it from Prelec et al.’s (2017) US States dataset, which we refer to as “States (PSM)”. As a baseline, we compared the performance of the SP algorithm to the average performance of all forecasters for



Please answer the questions below regarding this statement:

Birmingham is the capital of Alabama

Is this statement more likely to be true or false?

- ☐ True
☐ False

50 60 70 80 90 100

What is your estimated probability of being correct?

0 10 20 30 40 50 60 70 80 90 100

What percentage of other people do you think thought the bolded statement was true?

Imagine if the bolded statement was in fact **true**. What percentage of people would **correctly** think it was **true**?

Imagine if the bolded statement was in fact **false**. What percentage of people would **INCORRECTLY** think it was **true**?

>>

Figure 2.1: Example of a trial in Experiment 1.

each dataset, as well as the performance of standard algorithms in the literature using Cohen’s Kappa Coefficient, percentage accuracy.

2.4.2 Results and Discussion

Figures 2.2 and 2.3 show the performance for each algorithm on each of Prelec et al.’s (2017) datasets (left of dashed line) and our Experiment 1 dataset (right of dashed line). As shown in Table 2.2, the SP algorithm significantly outperformed all other algorithms on only two of the six datasets: the US States dataset reported by Prelec et al. (2017), and our Experiment 1 dataset which used the same set of US States questions. The SP algorithm did not significantly outperform all the other algorithms on the other four datasets.

These results suggest that the SP algorithm’s performance varies considerably across domains and only outperforms other algorithms for specific domains. To understand better how the algorithm operates, we conducted post-hoc analyses on these six datasets to understand how individuals’ performance varied as a function of their expertise.

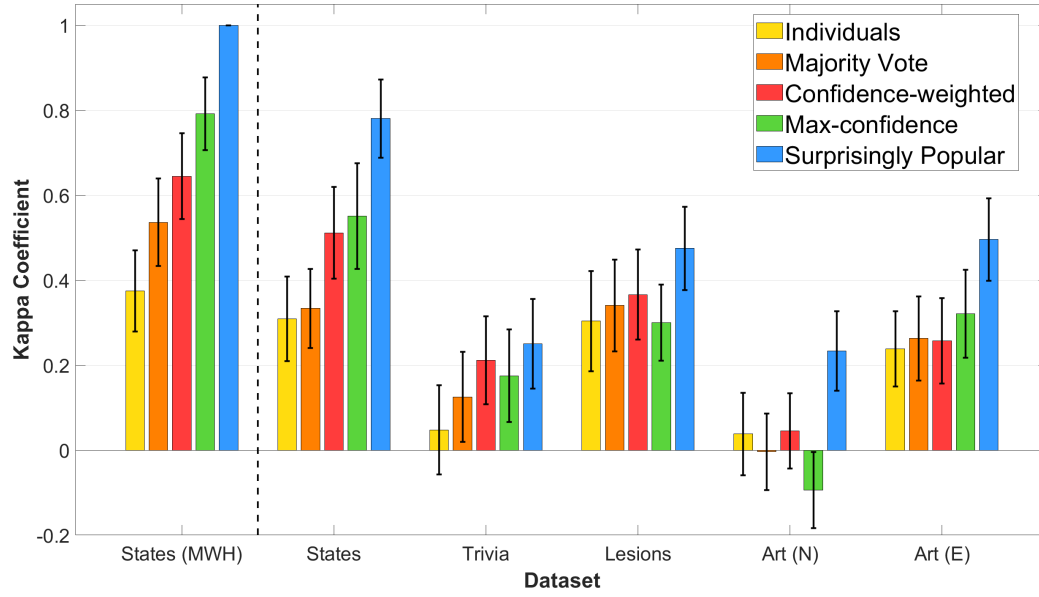


Figure 2.2: Classification performance of algorithms measured by Cohen’s Kappa Coefficient on each of Prelec et al.’s (2017) five datasets (left), and our dataset for Experiment 1 (right). Error bars show standard error.

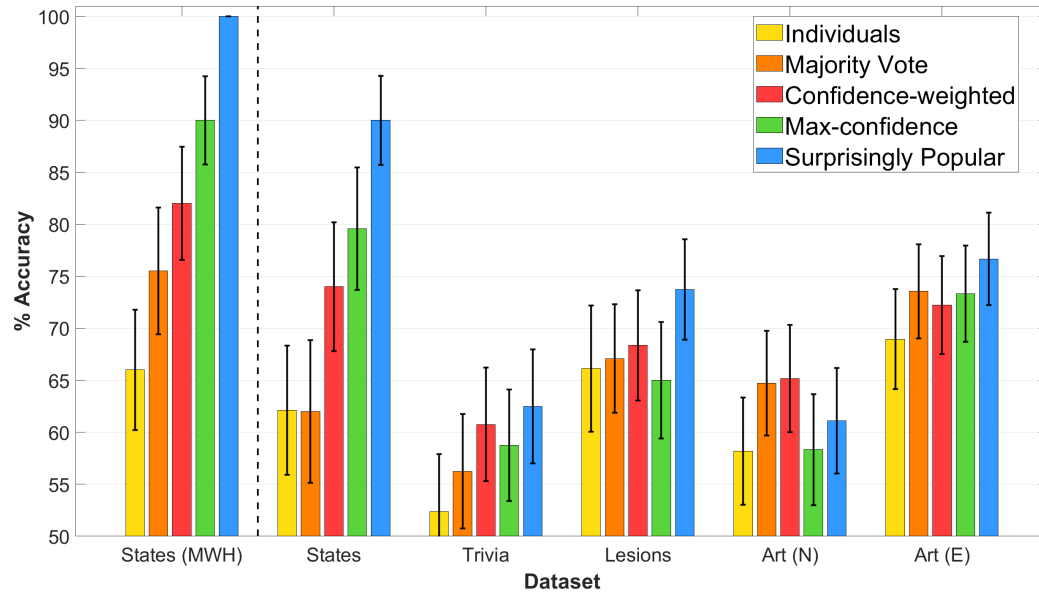


Figure 2.3: Classification performance of algorithms measured by percentage accuracy on each of Prelec et al.’s (2017) five datasets (left), and our dataset for Experiment 1 (right). Error bars show the standard error.

Table 2.2: Bootstrap 95% Confidence Intervals for Mean Paired Difference in Cohen’s Kappa Coefficient Between the SP Algorithm and Each Other Algorithm on Prelec et al.’s (2017) Datasets (Left of Dashed Line) and our Experiment 1 Dataset (Right of Dashed Line).

Algorithm	Dataset					
	States	Trivia	Lesions	Art (N)	Art (E)	States (MWH)
Majority Vote	[0.217, 0.657]*	[0.025, 0.245]*	[0.021, 0.262]*	[0.013, 0.443]*	[0.004, 0.451]*	[0.260, 0.662]*
Confidence-weighted	[0.037, 0.491]*	[-0.111, 0.185]	[-0.025, 0.251]	[-0.033, 0.396]	[0.018, 0.450]*	[0.163, 0.560]*
Max-confidence	[0.000, 0.484]	[-0.175, 0.325]	[-0.055, 0.387]	[0.102, 0.539]*	[-0.067, 0.419]	[0.044, 0.394]*

* indicates where the difference in Cohen’s Kappa Coefficient between the SP algorithm and the indicated algorithm was significant at the .05 level.

2.4.3 Analysis of expertise

We divided each of Prelec et al.’s (2017) five datasets and our Experiment 1 dataset into two halves using a median split, separating the high-performing (“High-performers”) and low-performing individuals (“Low-performers”) according to their percentage accuracy over all events. Figures 2.4 – 2.9 show the proportion of high-performers (red circles) and low-performers (blue crosses) voting “true” relative to their mean meta-predictions on each event, separated by whether the correct outcome was true (left panel) or false (right panel). For both plots, the horizontal (and vertical) distance from the reference line to each point corresponds to the absolute difference between forecasters’ votes and meta-predictions. As we showed in our reformulation of the SP algorithm, the SP algorithm weights individuals’ votes by the normalised absolute difference between individuals’ votes and meta-predictions. The distance between a particular point and the dotted line are therefore proportional to the weight given by the SP algorithm to that particular group for that event.

Recalling that in the original formulation of the SP algorithm, in order for the SP algorithm to predict the correct outcome for a particular group, the proportion of those forecasters voting “true” must be greater than their mean meta-prediction (i.e., the points must lie above the reference line) when the outcome is “true”, and conversely, the proportion of those forecasters voting “true” must be less than their mean meta-prediction (i.e., the points must lie below the reference line) when the outcome is “false”. The shaded regions in these plots therefore indicate where the SP algorithm would produce correct predictions.

Comparing the patterns in these plots across each dataset, we find that there are markedly different patterns of results across datasets. For the States (MWH) and States (PSM) datasets (Figures 2.4 and 2.5), we can see that there are clear differences between high-performing individuals’ and low-performing individuals’ patterns of responses. The left panel for these two datasets illustrates that for the events where the outcome is “true”, the proportion of high-performing individuals voting “true” was consistently lower than their mean meta-prediction. Similarly, the right panel for these two datasets illustrates that for the events where the outcome is “false”, the proportion of high-performing individuals voting “true” was consistently higher than their mean meta-prediction. What this indicates is that high-performing individuals were

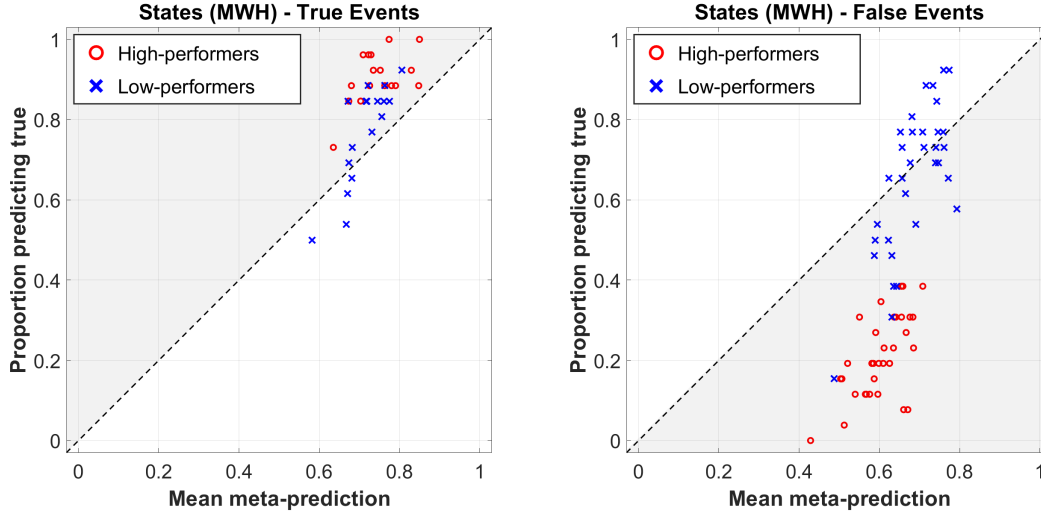


Figure 2.4: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the States (MWH) dataset. Each point represents that group’s average vote and meta-prediction for one event in the dataset. The diagonal line indicates where that group’s meta-predictions are exactly equal to their proportion of votes for “true”. The shaded regions indicate where the SP algorithm would generate correct predictions.

weighted very strongly by the SP algorithm relative to low-performing individuals, who tended to report meta-predictions approximately equal to the proportion of novices voting “true” (and therefore received weights close to 0).

In contrast, we see markedly different patterns of results for the remaining four datasets (Figures 2.6 to 2.9). The blue and red points on each plot are heavily overlapped in almost all four datasets, indicating that high-performing individuals’ and low-performing individuals’ responses are much more similar in terms of both votes and meta-predictions. Furthermore, both sets of points on all four plots are clustered around the reference line, which suggest that these differences in forecasters votes and mean meta-predictions are likely to reflect random noise rather than genuine differences in expertise. The SP algorithm generates the most accurate predictions when it is able to assign greater weights to high-performing forecasters than low-performing forecasters. As the differences between forecasters votes and meta-predictions in these datasets are ineffective for distinguishing between high-performing and low-performing forecasters, the SP algorithm would therefore not be expected to outperform other forecast-aggregation algorithms on these

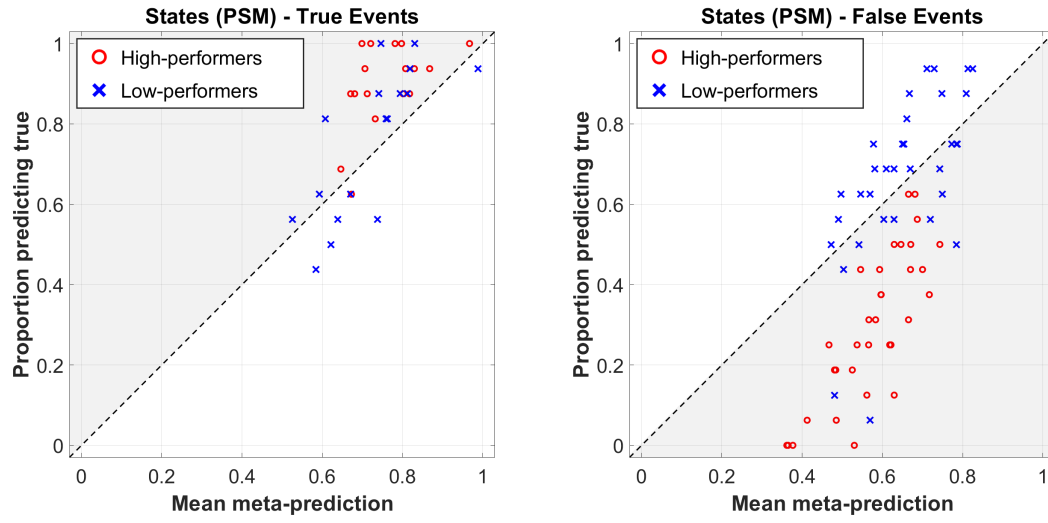


Figure 2.5: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the States (PSM) dataset.

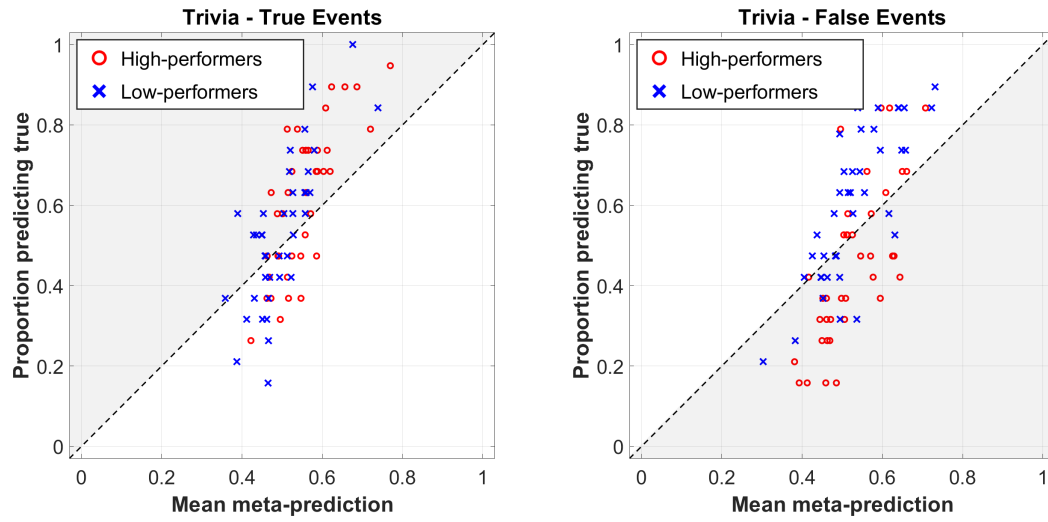


Figure 2.6: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Trivia dataset.

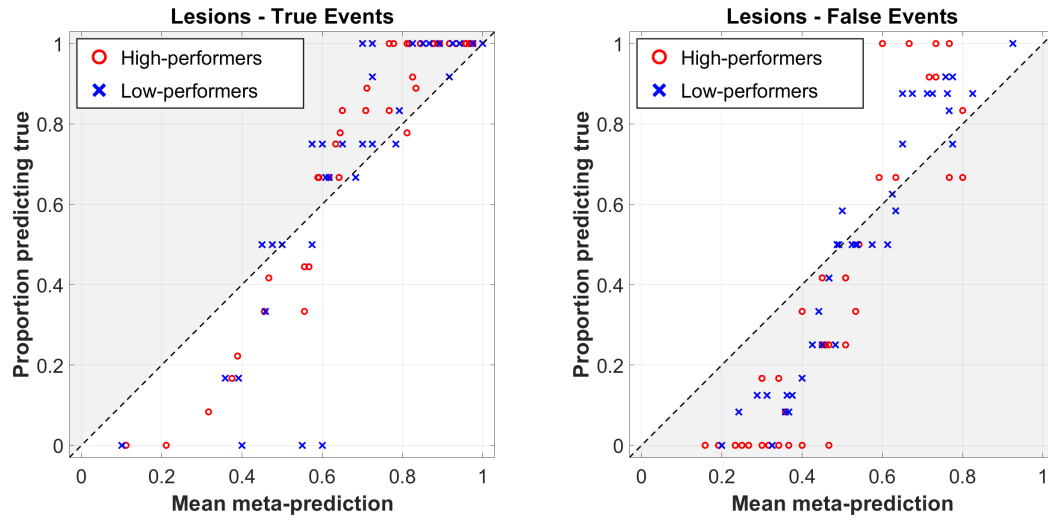


Figure 2.7: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Lesions dataset.

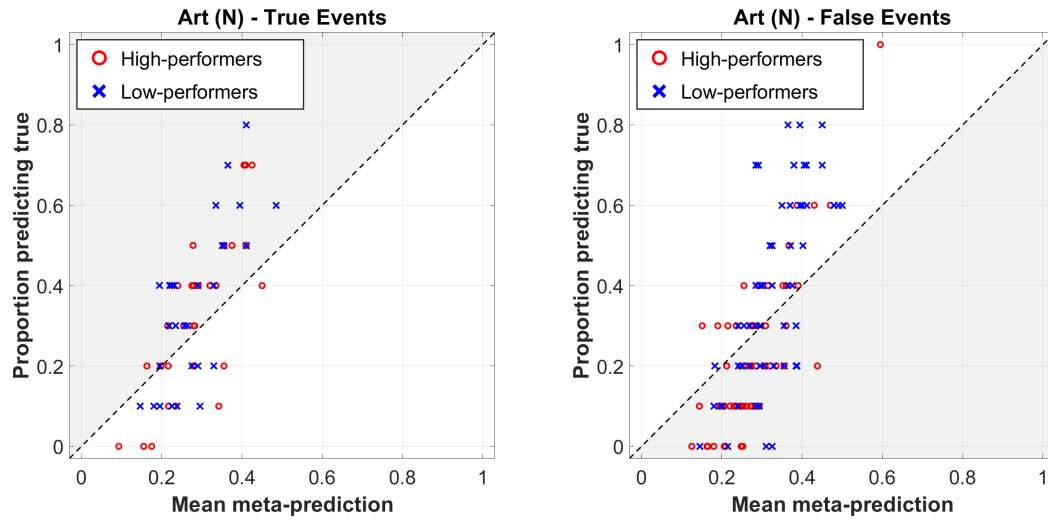


Figure 2.8: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Art (N) dataset.

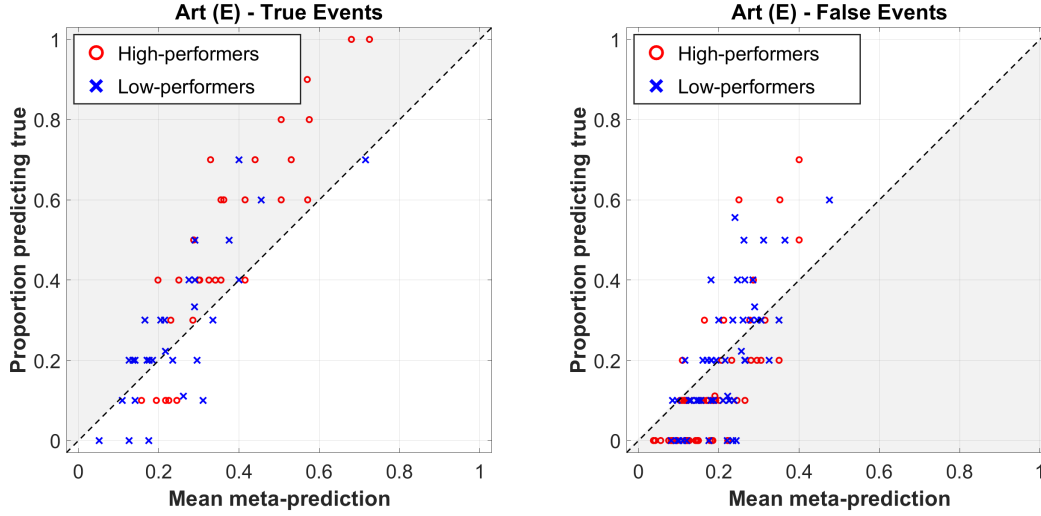


Figure 2.9: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions on the Art (E) dataset.

datasets. Indeed, if we re-inspect the SP algorithm’s performance on these four datasets in Figure 2.3, that seems to be case.

2.4.4 Identifying the SP weights assigned to individuals

We can better understand the SP mechanism by examining the weights assigned to forecasters by the SP algorithm in more detail. We ranked and sorted forecasters once again by their average accuracy across all the questions in a dataset. For each question, we binned forecasters into four quartiles based on their percentage accuracy and calculated the average weight assigned by the SP algorithm to each quartile. We then examined the weight assigned by the SP algorithm to forecasters in each of the four quartiles, averaged across all questions in the dataset.

Figure 2.10 plots the average weights assigned by the SP algorithm to the forecasters in each quartile, where forecasters have been sorted from lowest to highest accuracy, for each of the six datasets. For each plot, ‘Q1’ therefore contains the worst-performing forecasters and ‘Q4’ contains the best-performing forecasters. We can see that for the States (MWH) data and States (PSM) data, the SP algorithm consistently weights top-performing forecasters more than the worst-performing forecasters. Comparing the weights for the highest and lowest quartiles, we can

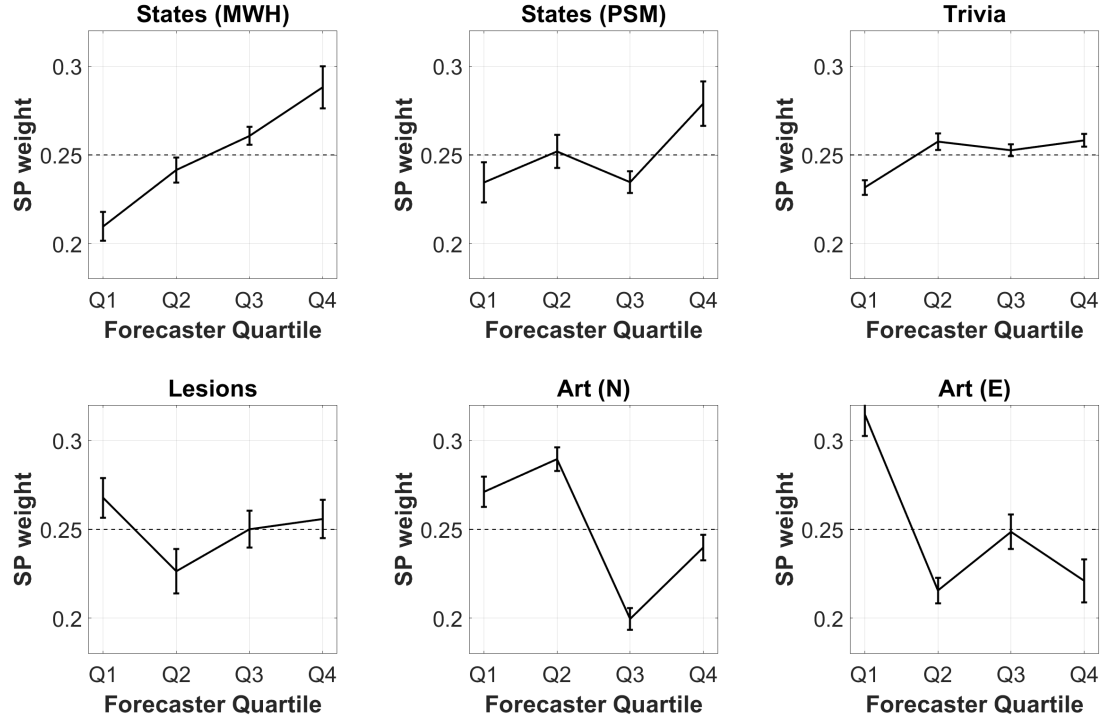


Figure 2.10: The average weight assigned by the SP algorithm as a function of forecasters' accuracy for each of the six datasets analysed in Experiment 1. Forecasters were sorted into four bins, with 'Q1' containing the least accurate forecasters and 'Q4' containing the most accurate forecasters. Error bars represent the standard error.

see that forecasters in the highest quartile were weighted almost twice as much as those in the lowest quartile for both the 'States' datasets. In contrast, the worst-performing forecasters in the other datasets were often weighted equal or more than the best forecasters. In particular, for both PSM's Art (novices) and Art (experts) datasets, we see that the best-performing forecasters were assigned lower weights than the worst-performing forecasters.

These results indicate that the SP algorithm's relatively poor performance on PSM's Lesions and Art (novices) datasets can be partly attributed to a failure to assign greater weights to the best-performing forecasters than the worst-performing forecasters in those datasets. The absolute differences between forecasters' votes and meta-predictions appear to be predictive of expertise in only the US States datasets but not the other datasets. In order to understand better why this is the case, we examine the distribution of forecasters' votes in each of these datasets.

2.4.5 Distribution of forecasters’ average accuracy

A possible explanation for why the SP algorithm does not significantly outperform other algorithms on Prelec et al.’s (2017) Trivia, Lesions, and Art (novices), and Art (experts) datasets is that the forecasters’ expertise in these datasets may be homogeneous, such that forecasters all have access to approximately the same quality information, and therefore there are no clear subset of experts in the crowd for the SP algorithm to identify.

To investigate this, we generated histograms plotting the distribution of forecasters’ percentage accuracy for each dataset, shown in Figure 2.11. The results demonstrate markedly different distributions of performance across datasets: For the two States datasets, there is clear evidence of heterogeneity in forecaster performance – both plots show strong evidence of multi-modality and high variability in forecaster performance. In contrast, the distributions in the other four datasets appear more uni-modal, and there appears to be much lower variability in forecaster performance.

In conjunction with Figures 2.10, these results suggest that the SP algorithm relies on heterogeneity in forecaster expertise in order for the differences between forecasters’ votes and meta-predictions to be predictive of expertise. When the crowd is homogeneous, differences in the weights assigned by the SP algorithm are likely to reflect noise rather than systematic individual differences in expertise. Majority voting, which is most accurate when expertise is homogeneous, is therefore likely to perform just as well as the SP algorithm under these conditions.

Altogether, the analyses in this section provide strong evidence that the SP mechanism relies on heterogeneous crowds with high variability in expertise in order to outperform other aggregation approaches. To identify the kinds of environments where such conditions are likely to be met, in the following section we use simulations to examine changes in the SP algorithm’s performance relative to that of other algorithms as we vary the level of crowd expertise.

2.4.6 Simulating changes in expertise

We simulated changes in crowd expertise for our US States dataset from Experiment 1 and compared the performance of the SP algorithm to the performance of the other algorithms to examine the range for which the SP algorithm was superior. We held meta-predictions constant

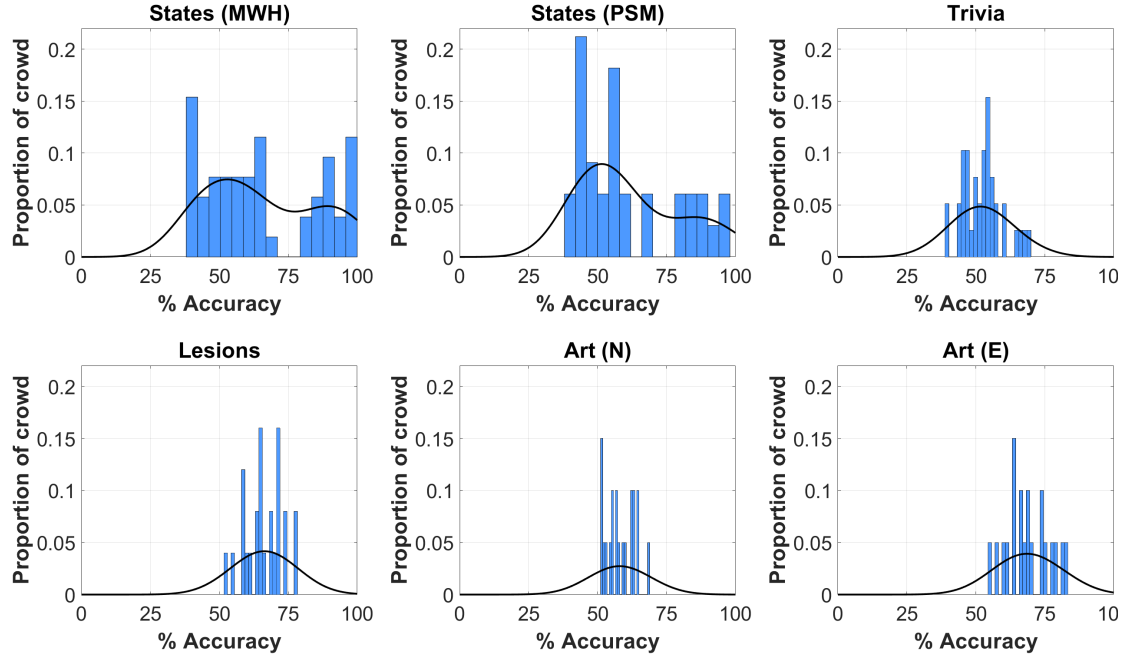


Figure 2.11: The distribution of forecasters over the percentage of events correctly predicted in each dataset analysed in Experiment 1.

while manipulating crowd expertise in order to provide a fair comparison across algorithms, as meta-predictions were only used by the SP algorithm. Manipulations of expertise were applied at the aggregate level and took one of three forms: adjustments to the aggregated crowd endorsement frequency (in the case of the majority vote and SP algorithms), the aggregated probability estimate (for the confidence-weighted algorithm), and difference between the aggregated confidence for those who predicted the correct answer and those who did not (in the case of the max-confidence algorithm). These aggregate values being adjusted thus correspond to the left-hand side of the inequalities for each algorithm in Table 2.1 when the correct answer is “true”, and the right-hand side when the correct answer is “false”. Thus, as the value on the side of the inequality favouring the correct answer increased, expertise increased.

Adjustments were made in .01 increments, starting from an adjustment of -.5 (i.e., a strong decrease in crowd expertise) to an adjustment of +.5 (i.e., a strong increase in crowd expertise). For example, an adjustment of +.5 meant an increase in the aggregate proportion of votes for the correct outcome of each question by .5, an increase in the average probabilistic forecast for the correct outcome by .5, or an increase in the difference of average confidence for the correct

answer by .5. We generated 1,000 bootstrap samples, applied the increment/decrement to each bootstrap sample, evaluated the equation for each algorithm to generate a prediction, and then evaluated the performance of each algorithm. The top panels of Figures 2.12–2.13 show the change in performance for each of the four aggregation algorithms as expertise is adjusted. The bottom panels of each figure show the difference in performance between the SP algorithm and each other algorithm as expertise is adjusted.

Across all four metrics of performance, these simulations demonstrate that the SP algorithm is highly sensitive to the level of expertise in the crowd, and that there is a clear non-monotonic relationship between the performance of the SP algorithm and the performance of each of the other algorithms. At moderate levels of expertise (around -.1 adjustment), the SP algorithm improves at a faster rate than other algorithms as expertise increases, but its performance also decreases more rapidly than other algorithms as expertise decreases. There is a clear ceiling effect where all four algorithms reach the same level of performance at higher levels of expertise. In contrast, the SP algorithm is outperformed by both the majority vote and confidence-weighted algorithms at low levels of expertise. The SP algorithm is shown to perform best relative to all three other algorithms at moderate levels of expertise, where it appears to be making use of crowd expertise more efficiently than other algorithms.

In Wilkening et al. (2020), we show that difference between the total weight given to experts and the total weight given to novices is maximised when there is a moderate number of experts – an environment where heterogeneity in expertise is likely to be maximised. This suggests that, consistent with Figures 2.12–2.13, there may be a non-monotonic relationship between expertise and the performance of the SP algorithm relative to other algorithms, with the SP algorithm having the best relative performance when there is a moderate number of experts.

In the next section, we report the results from an experiment where we test for the non-monotonic relationship between expertise and the performance of the SP algorithm relative to other algorithms in a domain where we can induce systematic variation in expertise that is naturally ordered.

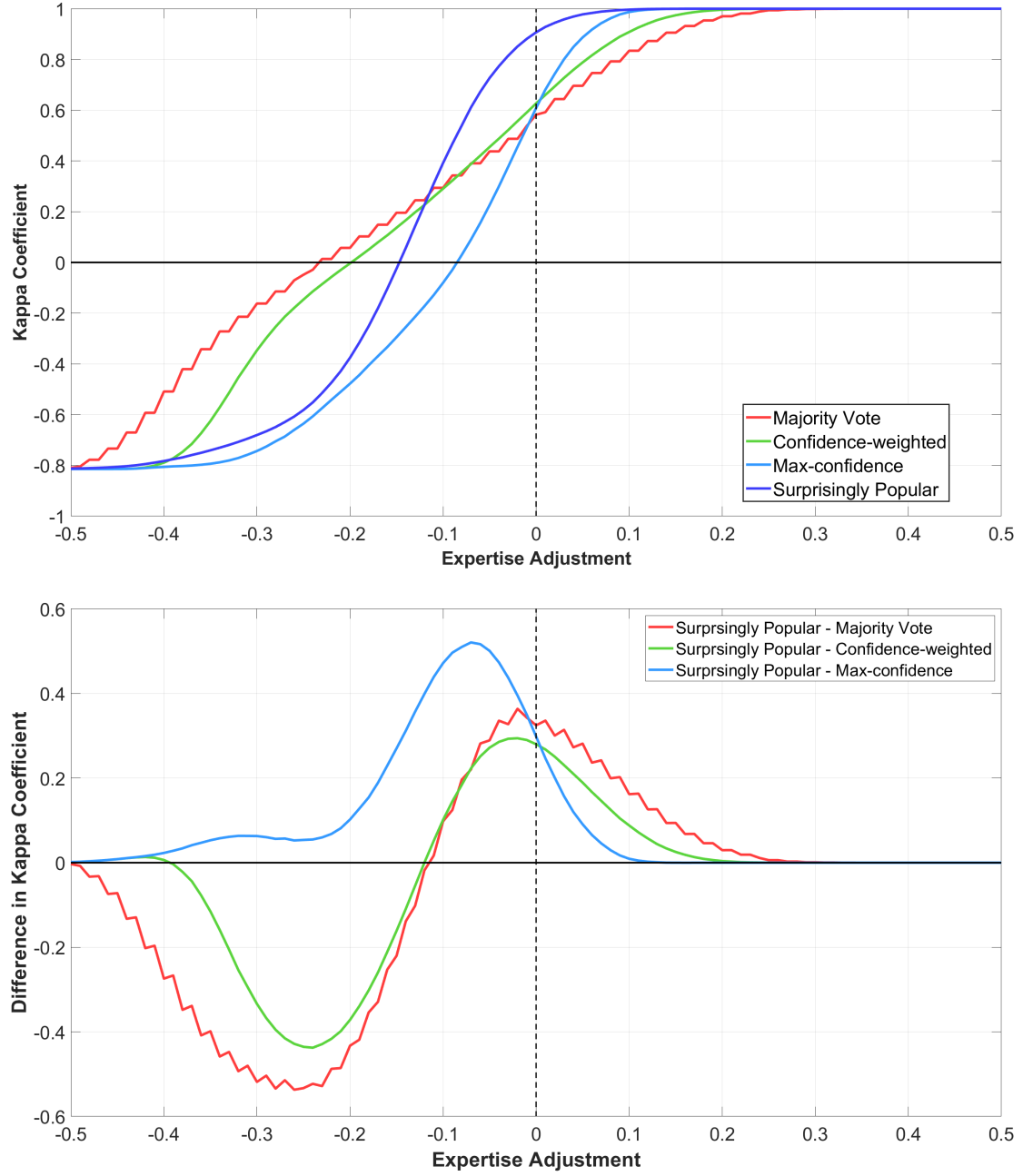


Figure 2.12: **Top panel:** The change in Cohen's Kappa Coefficient of each algorithm as crowd expertise is varied in the States (MWH) dataset from Experiment 1. The vertical dashed line indicates the performance of each algorithm without any expertise adjustment. **Bottom panel:** The difference in Cohen's Kappa Coefficient between the SP algorithm and each other algorithm as a function of different adjustments to expertise.

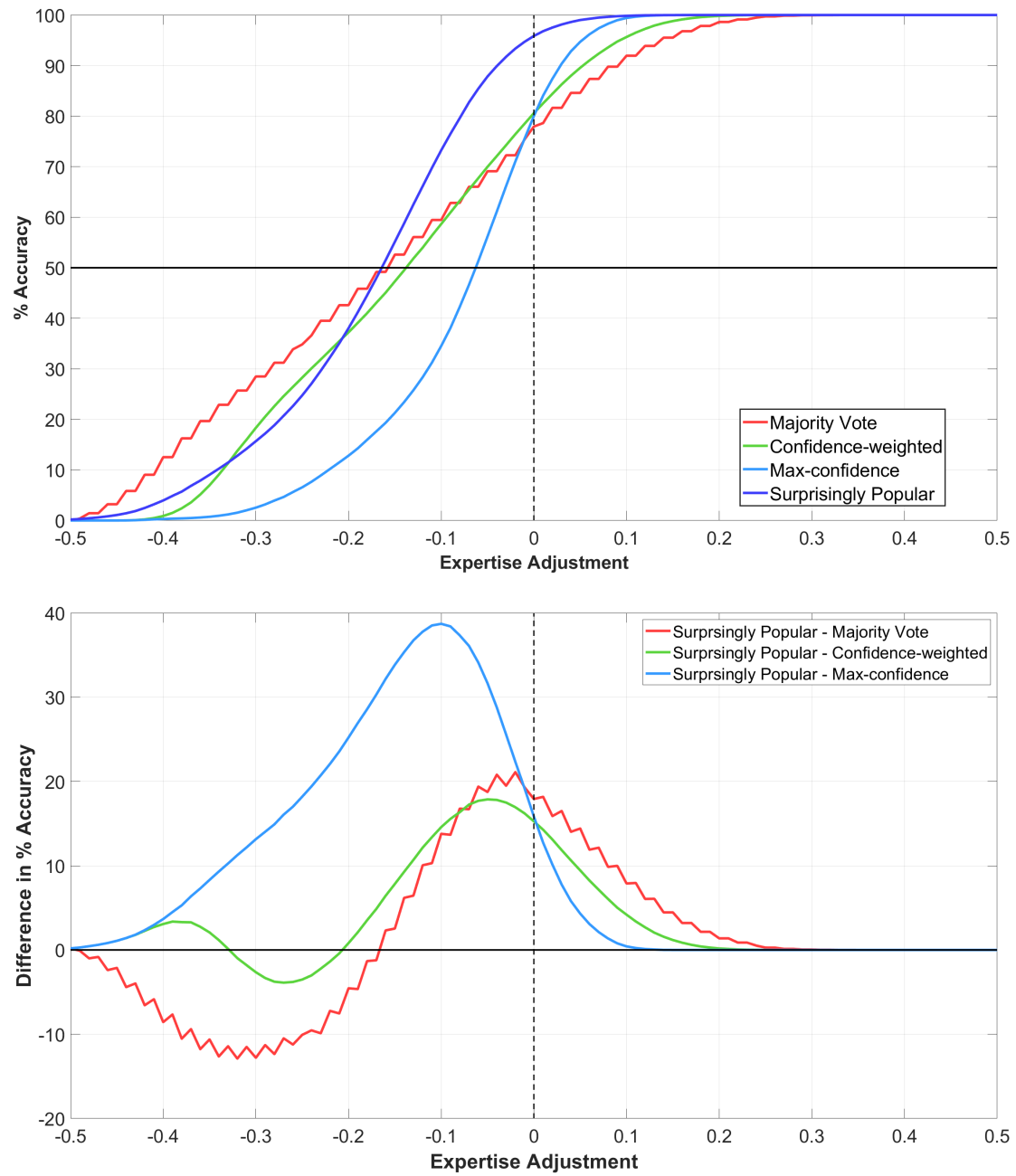


Figure 2.13: **Top panel:** The change in percentage accuracy of each algorithm as crowd expertise is varied in the States (MWH) dataset from Experiment 1. The vertical dashed line indicates the performance of each algorithm without any expertise adjustment. **Bottom panel:** The difference in percentage accuracy between the SP algorithm and each other algorithm as a function of different adjustments to expertise.

2.5 Experiment 2

The aim of Experiment 2 is to provide a rigorous test of the non-monotonic relationship between expertise and SP performance proposed in the previous section. We examine the SP algorithm’s performance relative to the performance of other algorithms as the amount of expertise in the crowd is systematically varied. To vary the amount of expertise in the crowd, we manipulated task difficulty with the assumption that crowd expertise decreased as task difficulty increased. In line with our results above, we predict that the SP algorithm outperform majority weighting for questions of moderate difficulty (i.e., difficulties 2–4), but would offer no significant improvement over competing algorithms for questions at the hardest (5) and easiest difficulty (1). Furthermore, upon conducting the same analyses from the preceding section on the dataset at each of these difficulties, we would expect to see a similar pattern of results for the moderate difficulty compared to what we observed in Figures 2.4–2.11 for the US States datasets, but not for the other difficulties. Overall, findings consistent with these predictions would provide strong evidence that a fundamental mechanism of the SP algorithm is to identify and leverage latent expertise in the crowd.

2.5.1 Methods

We generated 500 science statements at a US primary and secondary grade school level. Questions were adapted from worksheets on the Education Quizzes website (<http://www.educationquizzes.com/us>), and then converted into true or false statements. Approximately 2-3 questions were taken from each worksheet from the Biology, Chemistry, Geography, Physics, and General Science categories, spanning from grades 1 to 12, broken up into five levels of difficulty (grades 1 and 2; grades 3, 4, and 5; grades 6, 7, and 8; grades 9 and 10; and grades 11 and 12). We coded “Difficulty 1” as the easiest difficulty, and “Difficulty 5” as the hardest difficulty. We treated each set of 100 questions of the same difficulty as an individual dataset. The list of questions in our experiment are included in the Appendices section (see Section 6.1).

We recruited 500 respondents from Amazon Mechanical Turk; only respondents inside the US were able to participate in the experiment. Participants were paid a flat fee of \$4.00 for

completing the survey. The survey was conducted on the Qualtrics platform. Participants were asked to answer each question as honestly as they could, and were asked not to cheat (e.g., by looking up any of the questions online). Six individuals who reported cheating at the task or had failed to complete the survey were excluded from the analyses; analyses were conducted on the data of the remaining 494 participants.

Participants completed 100 trials each, with each trial comprising one statement which was either true or false, and then followed by the same three questions we asked in Experiment 1 (see Section 2.4). An example of a trial is shown in Figure 2.14. Half the statements at each level of difficulty were true, and the other half were false. Each participant saw 20 statements from each level of difficulty, and statements were presented in one of five randomised orders. Participants who took part in any of our previous experiments were excluded from participating. Data collection for all five datasets was completed in August 2017.

2.5.2 Results

The mean performance of each algorithm at each level of difficulty is shown in Figures 2.15–2.16. Results showed a monotonic decrease in mean individual performance as questions became harder, indicating that the average level of expertise in the crowd had decreased as questions became harder and thus our difficulty manipulation was effective. Nonetheless, there was very little difference between algorithms’ performance on the two hardest difficulties, indicating that difficulty had increased much more quickly than we had anticipated.

Table 2.3 shows the bootstrap 95% CIs for paired mean difference in Cohen’s Kappa between the SP algorithm and other algorithms, for each difficulty. The SP algorithm outperformed all other algorithms on questions of moderate difficulty (Difficulty 3), although only the difference in performance between the SP algorithm and majority voting was significant. There was no significant difference in performance between the SP algorithm and other algorithms on any of the other difficulties. All four algorithms appear to have extremely similar levels of performance at each level of difficulty, suggesting that the advantage of the SP algorithm over other algorithms may be confined to a narrower range of moderate difficulty problems (and therefore environments where expertise is even more heterogeneous) than we had originally expected.



Faculty of Medicine, Dentistry and Health Sciences
Psychological Sciences

6: Amphibians are warm blooded.

Is this statement more likely to be true or false?

☐ True
☐ False

5060708090100

What is your estimated probability of being correct?

0102030405060708090100

What percentage of other people do you think thought the bolded statement was true?

>>

Figure 2.14: Example of a trial in Experiment 2.

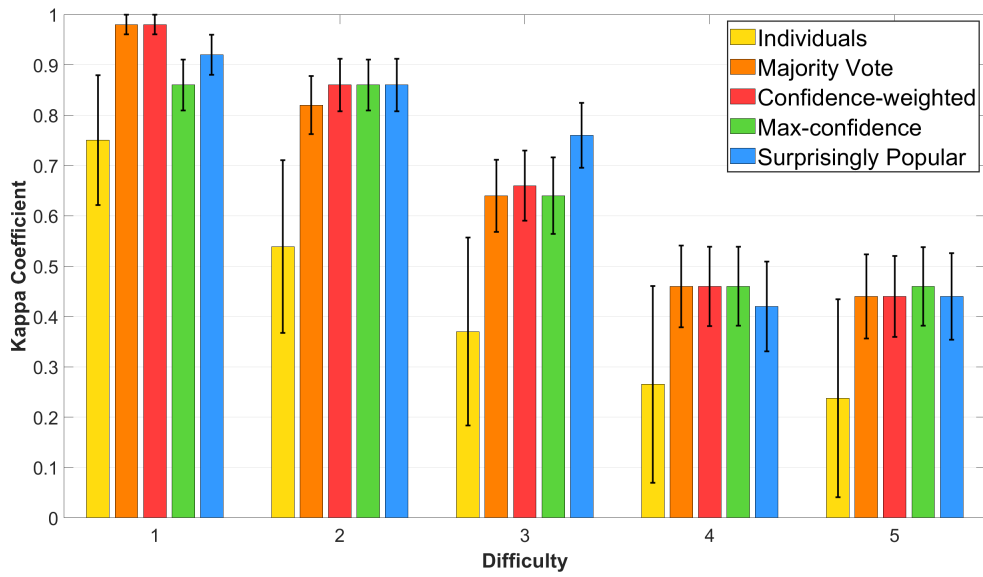


Figure 2.15: The mean Cohen's Kappa Coefficient and standard error for each algorithm across each level of question difficulty (1 - easiest to 5 - hardest).

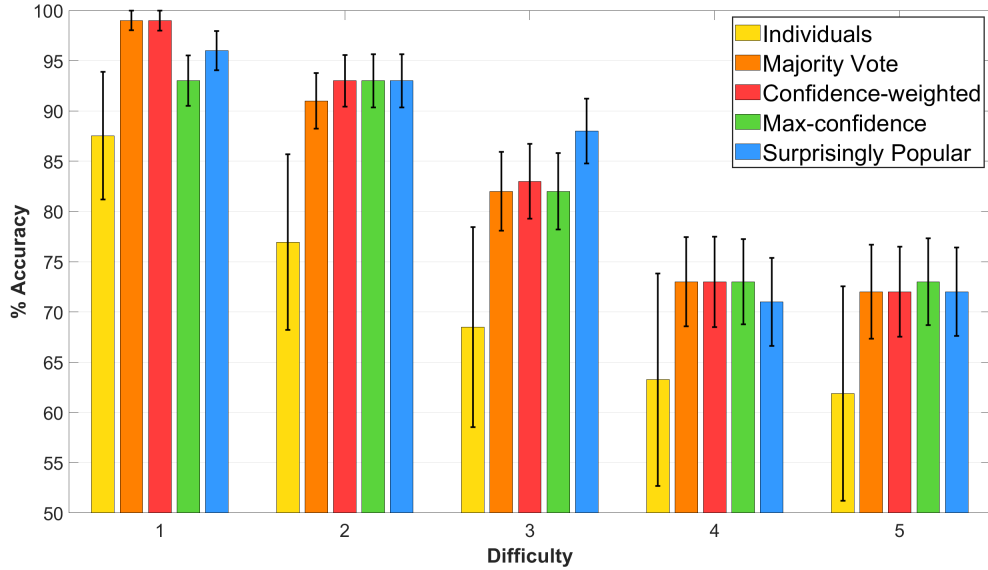


Figure 2.16: The mean percentage accuracy and standard error for each algorithm across each level of question difficulty (1 - easiest to 5 - hardest).

Table 2.3: Bootstrap 95% Confidence Intervals for Mean Paired Difference in Cohen's Kappa Coefficient Between the SP Algorithm and Each Other Algorithm Across Different Levels of Difficulty.

Algorithm	Difficulty				
	1 - Easiest	2	3	4	5 - Hardest
Majority Vote	[-0.139, 0.000]	[-0.040, 0.119]	[0.021, 0.229]*	[-0.121, 0.036]	[-0.080, 0.078]
Confidence-weighted	[-0.139, 0.000]	[-0.100, 0.098]	[-0.013, 0.215]	[-0.139, 0.052]	[-0.098, 0.092]
Max-confidence	[-0.059, 0.181]	[-0.120, 0.106]	[-0.036, 0.273]	[-0.193, 0.102]	[-0.230, 0.184]

* indicates where the difference in Cohen's Kappa Coefficient between the SP algorithm and the indicated algorithm was significant at the .05 level.

2.5.3 Analysis of expertise

Figures 2.17–2.23 provide a more rigorous exploration of these results. Figures 2.17–2.21 show that the pattern of high-performing individuals’ (“High-performers”) and low-performing individuals’ (“Low-performers”) proportion of “true” votes to mean meta-predictions varied systematically across each difficulty.

For Difficulties 1 to 2 (i.e., the two easiest difficulties), high-performing individuals’ and low-performing individuals’ predictions were highly intermixed, indicating that there was little difference between both groups’ patterns of response. As difficulty increased, low-performing individuals’ points became increasingly clustered towards the diagonal line, indicating that the difference between low-performing individuals’ votes and meta-predictions moved further towards zero as difficulty increased. For difficulties 4 and 5, high-performing individuals’ responses had also migrated towards the diagonal line and similar to the easiest difficulties, and there was little difference between high-performing individuals’ and low-performing individuals’ patterns of response.

In contrast, for Difficulty 3, we can see that high-performing individuals exhibited notably different patterns of response to low-performing individuals. High-performing individuals’ responses are primarily located in the shaded region of each plot, whereas low-performing individuals’ responses are noisily clustered around the diagonal reference line. This was not the case for the other difficulties, where both groups’ points were much more intermixed. The weights assigned by the SP algorithm, which are determined by these absolute differences, can therefore discriminate between high-performing individuals and low-performing individuals most effectively on questions of moderate difficulty.

As we had predicted, the pattern of responses for Difficulty 3 (Figure 2.19) was most similar to the responses in the US states datasets from Prelec et al. (2017) and our Experiment 1 (Figures 2.5 and 2.4), which suggests that the questions of moderate difficulty had similar properties to the two US States datasets. Nonetheless, high-performing individuals’ and low-performing individuals’ points in Figure 2.19 were not as well-separated as those we observed for either US States datasets, suggesting that there was not the same level of heterogeneity of expertise present in the Difficulty 3 dataset compared to either of the US states datasets.

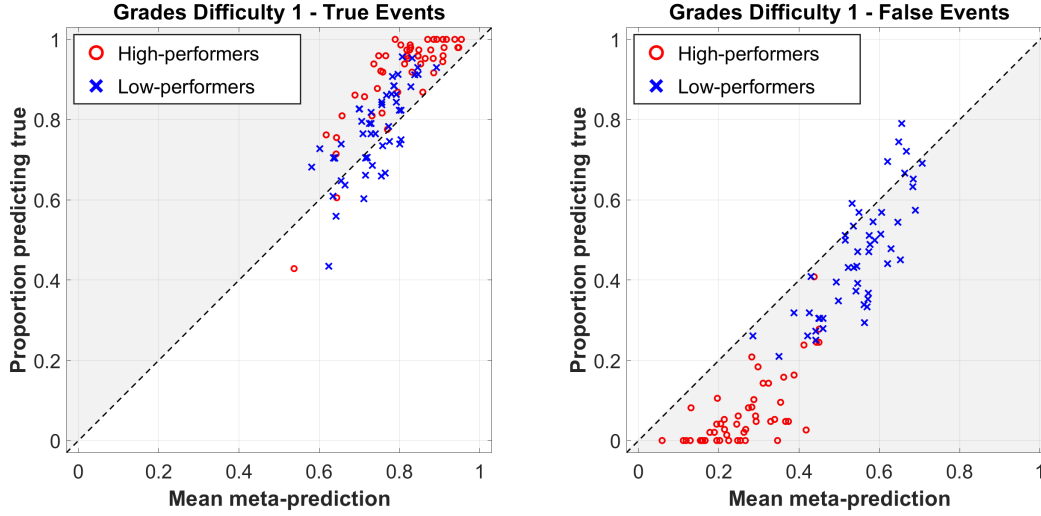


Figure 2.17: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions, for our Experiment 2 US Grades Difficulty 1 dataset. Each point represents that group’s average vote and meta-prediction for one event in the dataset. The diagonal line indicates where that group’s meta-predictions are exactly equal to their proportion of votes for “true”. The shaded regions indicate where the SP algorithm would generate correct predictions.

In general, across all five difficulties, forecasters’ responses demonstrated lower variability for the questions where the correct outcome was “true” compared to questions where the correct outcome was “false”. This result suggests that, similar to what we had observed in the US states datasets, questions where the outcome was “false” was most effective for discriminating differences in forecasters’ expertise.

2.5.4 Weights assigned by the SP algorithm

Comparing the weights assigned by the SP algorithm to each forecaster quartile in Figure 2.22, we can see that the SP algorithm assigns different patterns of weights for the easiest difficulties (1 and 2) compared to the other three difficulties. For the easiest two difficulties, the SP algorithm incorrectly assigns the highest weights to the worst quartiles, which explains why it performs slightly (but not significantly) worse than majority voting on the easiest difficulty. Surprisingly, on the moderate to hardest difficulties (3, 4, and 5), the SP algorithm assigned approximately

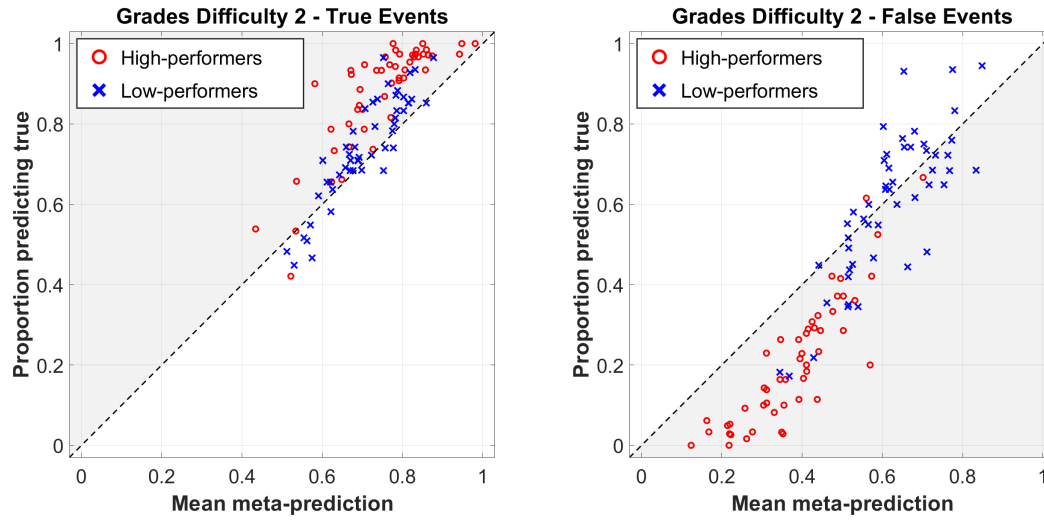


Figure 2.18: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions, for our Experiment 2 US Grades Difficulty 2 dataset.

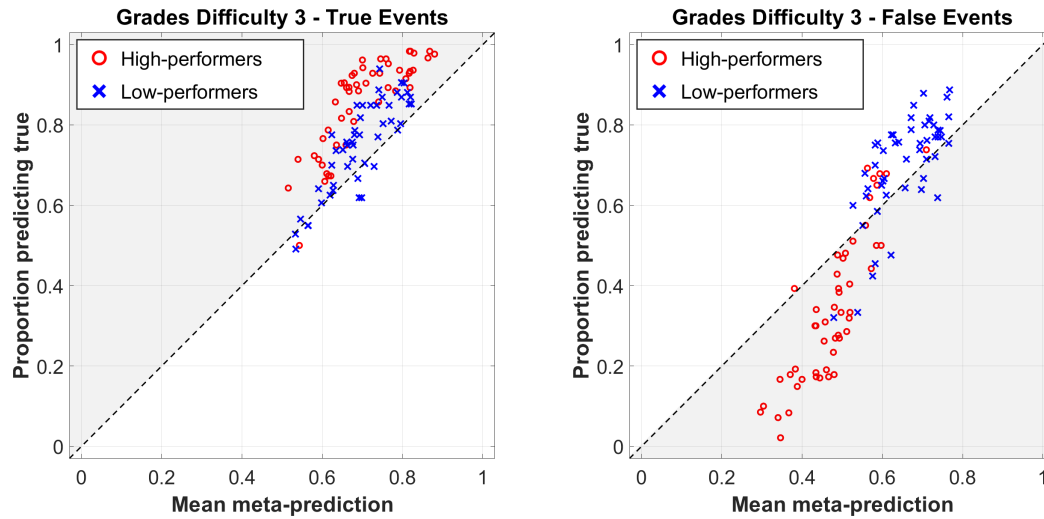


Figure 2.19: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions, for our Experiment 2 US Grades Difficulty 3 dataset.

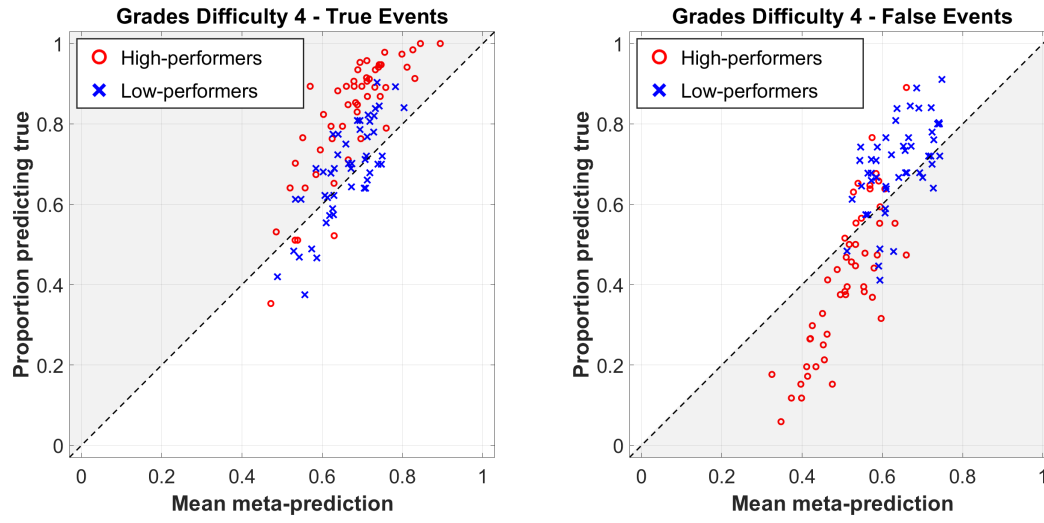


Figure 2.20: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions, for our Experiment 2 US Grades Difficulty 4 dataset.

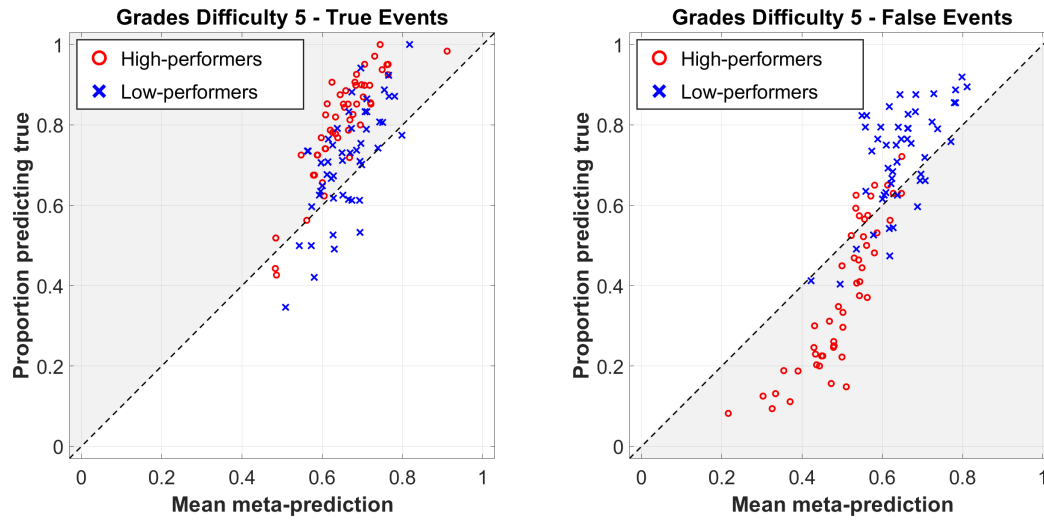


Figure 2.21: Mean proportion of high-performing individuals and low-performing individuals predicting events to occur where the outcome was true (left panel) and false (right panel), compared to their mean meta-predictions, for our Experiment 2 US Grades Difficulty 5 dataset.

equal weight to each of the four quartiles, and was therefore, on average, unable to identify the high-performing individuals in the crowd. Despite this, the SP algorithm still significantly outperformed majority voting on Difficulty 3 (i.e., questions of moderate difficulty). This suggests that the SP algorithm had therefore (1) correctly assigned greater weights to high-performing individuals on questions where the SP algorithm voted against the majority, and (2) over-weighted low-performing individuals on questions where the SP algorithm voted with the majority.

Indeed, when we calculate the average weight assigned by the SP algorithm to forecasters on problems where the SP algorithm voted against the majority of forecasters (8 out of 100 questions), the best-performing quartile of forecasters receive an average SP weight of .277, whereas the worst-performing quartile of forecasters receive an average SP weight of .207. Thus, on questions where the SP algorithm disagrees with majority vote, the SP algorithm was able to identify the expertise of high-performing individuals and assign weights accordingly.

In contrast, on questions where both algorithms concurred (92 out of 100 of questions), the best-performing quartile of forecasters was assigned an average weight of .247 whereas the worst-performing quartile of forecasters was assigned an average weight of .262. Thus, although the SP algorithm over-weighted low-performing individuals by a small amount on the vast majority of questions, it was able to generate significantly better predictions than majority voting overall because it could identify high-performing individuals on a subset of questions in the dataset.

2.5.5 Distribution of forecasters' percentage accuracy

Comparing the distributions of forecasters' accuracy for each of these five datasets in Figure 2.23, we can see that the distribution for Difficulty 3, where the SP algorithm outperformed majority voting, has the greatest variability out of all five difficulties. This provides further evidence that the SP algorithm indeed performs best in high-variability environments. Nonetheless, there appears to be high uni-modality in the distribution of scores of all five difficulties, indicating that there are no strong heterogeneous differences in expertise in any of the five datasets. Thus, our experimental manipulation did not successfully induce distinct sub-groups of high-performing individuals and low-performing individuals in the crowd. This explains why the improvement offered by the SP algorithm over other aggregation approaches for the Difficulty 3 dataset was

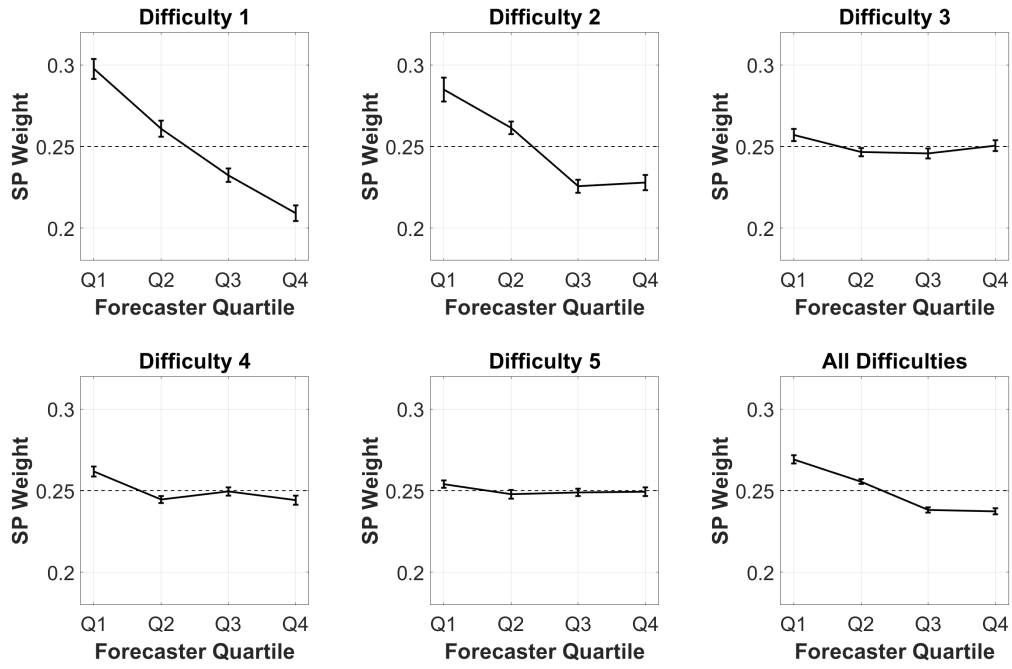


Figure 2.22: The average weight assigned by the SP algorithm as a function of forecasters' accuracy in each of the five US Grades datasets in Experiment 2. Forecasters were sorted into four bins, with 'Q1' containing the least accurate forecasters and 'Q4' containing the most accurate forecasters. Error bars represent the standard error.

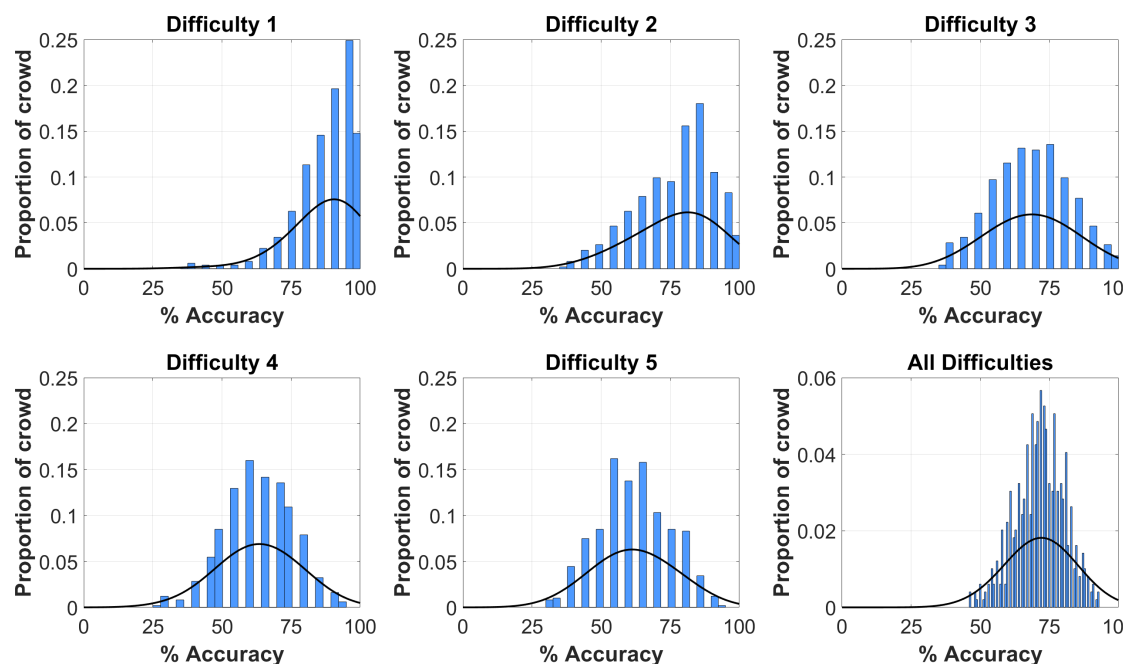


Figure 2.23: The distribution of forecasters over the percentage of events correctly predicted in each of the five datasets from Experiment 2.

much smaller than what we had observed in the US states datasets.

2.5.6 Discussion

As we had predicted, the SP algorithm offered the greatest improvement on questions of moderate difficulty. Consistent with our results in Experiment 1 (Figures 2.15–2.11), we see that the SP algorithm’s performance relative to other algorithms is maximised when variability in forecasters’ expertise is highest (Figure 2.23). These results therefore provide further evidence that the performance of the SP algorithm relative to other algorithms has a non-monotonic relationship to the level of expertise in the crowd. When crowds contain distinct subsets of experts and non-experts, the SP algorithm is able to leverage the experts’ votes using the absolute difference between their vote and meta-prediction. In contrast, when the crowd is homogeneous, differences in the weights assigned by the SP algorithm is most likely due to random noise rather than meaningful differences in expertise.

2.6 Robustness Over Different Sample Sizes

Additionally, we investigated the robustness of the SP algorithm’s performance over different sample sizes for each of the datasets in Experiment 2. While both Prelec et al.’s (2017) model and our theoretical model predict the SP algorithm’s performance to increase and converge as a function of sample size, the interaction between sample size and difficulty remains to be explored empirically. Although Prelec et al. (2017) showed that the SP algorithm was robust with crowds as small as 20 people, it is not clear whether this robustness extends to even smaller-sized crowds, or whether this is the case for different types of forecasting problems.

We used bootstrap resampling to simulate changes in performance for each algorithm over different sample sizes, for each difficulty dataset in Experiment 2. For each event and each bootstrap sample size n from 10 to 100 forecasters in increments of 10, we randomly resampled n forecasters from the original sample with replacement. For each event, we only sampled from forecasters who made a prediction for that event. On each of these bootstrap samples, we aggregated forecasters’ responses using the majority vote, confidence-weighted, max-confidence, and the SP algorithm. We repeated this 1,000 times for each event and each n , computed Cohen’s Kappa Coefficient for each algorithm, and plotted the change in performance as a function of sample size.

Figure 2.24 shows the change in performance for each algorithm over different sample sizes on each of the five US grades datasets. For questions of moderate difficulty (i.e., Difficulty 3), the performance of the SP algorithm was highly robust even for smaller sized samples with at least 25 forecasters in the crowd. The SP algorithm’s performance decreased at a similar rate for the other difficulties, thus, its robustness to small sample sizes therefore does not depend on the algorithm’s performance relative to other aggregation approaches. These simulation results are consistent with results observed for Prelec et al.’s (2017) datasets, which used relatively smaller sample sizes than our datasets and found that the SP algorithm still performed well.

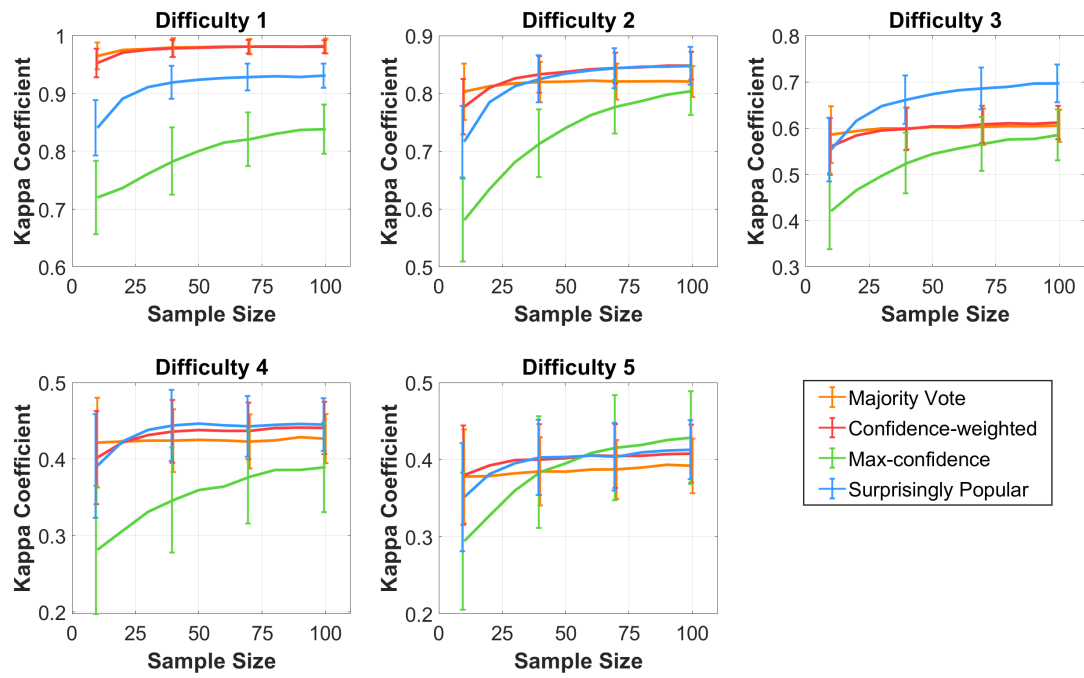


Figure 2.24: Simulation results showing the change in algorithms' performance in terms of Cohen's Kappa Coefficient over different sample sizes for each of the five datasets from Experiment 2. Error bars show the standard error.

2.7 Testing the SP Mechanism on NFL Predictions

As another test of the proposed SP mechanism, we repeated our analyses using the data collected by Lee et al. (2018). Lee et al. (2018) applied the SP algorithm on predictions on the outcomes of 2017-2018 US National Football League (NFL) games. While the SP algorithm has been applied across a range of domains and decision making situations, its performance had yet to be compared to standard forecast-aggregation approaches on questions of genuine predictions (i.e., about the outcomes of future events). Lee et al. (2018) found that out of 256 NFL games over a period of 17 weeks, using a crowd of forecasters who rated themselves as ‘extremely knowledgeable’ in NFL, the SP algorithm performed slightly worse than taking the majority vote from that crowd of forecasters. Similarly, the SP algorithm, when applied to the crowd of ‘extremely knowledgeable’ forecasters, performed slightly worse than the majority vote of media experts.

As the data from this study was kindly shared by the authors, the current section looks to re-analyse their results in light of our findings from Experiments 1 and 2 and examine whether the SP algorithm performs as expected under environments with heterogeneity and high variance in expertise as we would expect. While Lee et al. (2018) found that the SP algorithm offered no significant improvement over other algorithms at the aggregate level, finer analysis of the performance of the SP algorithm at the event level could potentially provide useful insight into the SP algorithm’s mechanisms. We therefore generated predictions for the questions in their dataset using the same algorithms from Experiments 1 and 2 above and compared the predictions of each algorithm. Applying the same set of analyses as before, we therefore examined: (1) the differences in the patterns of response between high-performing and low-performing individuals in the crowd, (2) the weights assigned by the SP algorithm to each quartile of forecasters, (3) the distribution of forecasters’ accuracy.⁵

Figure 2.25 shows the percentage accuracy of each algorithm each week. We can see that the SP algorithm, majority voting, the confidence-weighted algorithm, and the max-confidence algorithm indeed provide very similar predictions each week. The SP algorithm (blue bar) offers

⁵We note that the NFL dataset has an interesting property where a group of different participants provided predictions over each week. Interestingly, Experiment 2’s US Grades dataset also had a similar property, where each forecaster only answer 100 out of the 500 questions in the Experiment, depending on which set out of the five questions sets that forecaster was allocated.

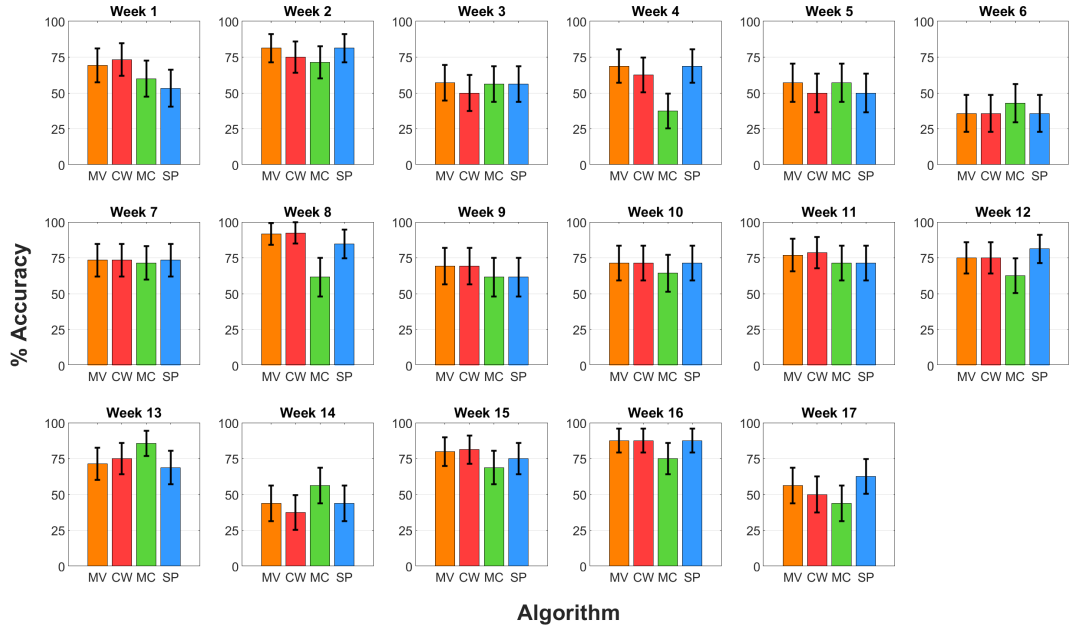


Figure 2.25: Percentage accuracy for each algorithm for each week from Lee et al.’s (2018) NFL dataset.

a very small improvement in accuracy for weeks 12 and 17, but otherwise does not outperform other algorithms for any of the other weeks in this dataset.

Figure 2.26 shows the differences between high-performing individuals’ and low-performing individuals’ aggregate patterns of votes and meta-predictions each week. The large vertical spread of points for each week is indicative of the variability of the difficulty of predictions each week. However, for most weeks, high-performing individuals and low-performing individuals appear to exhibit similar patterns of responses (i.e., the red circles and blue crosses are consistently overlapping). Forecasters therefore appear to be homogeneous in expertise in the majority of weeks.

Examining Figure 2.27, which shows the weights assigned by the SP algorithm to the best-performing and worst-performing forecasters for each week, we can see that the SP algorithm is very inconsistent from week-to-week in assigning greater weights to the top quartile of forecasters. This includes weeks 12 and 17, where the SP algorithm offers minor improvements over other algorithms. The SP algorithm assigns greater weights to the best-performing individuals in week 17 but shows the opposite pattern in week 12. The absolute difference between forecasters’ votes

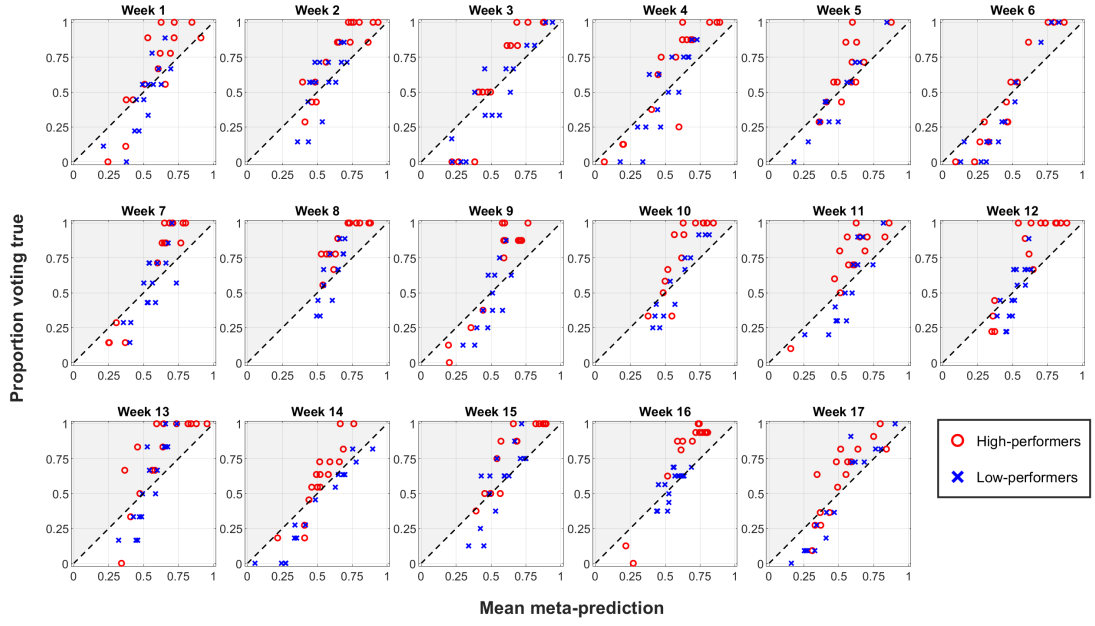


Figure 2.26: Mean proportion of low-performing individuals (blue crosses) and high-performing individuals (red circles) predicting events to occur compared to their mean meta-predictions for Lee et al.’s (2018) NFL dataset. Each point represents that group’s average vote and meta-prediction for one event in the dataset. The diagonal line indicates where that group’s meta-predictions are exactly equal to their proportion of votes for “true”. The shaded regions indicate where the SP algorithm would generate correct predictions.

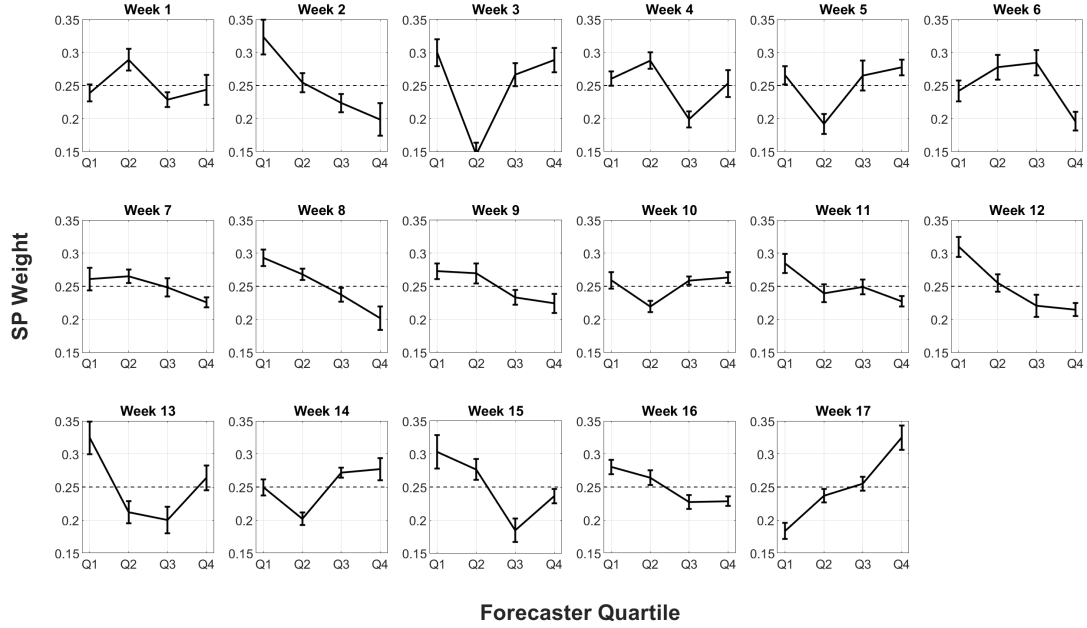


Figure 2.27: The average weight assigned by the SP algorithm to the votes of forecasters as a function of the performance of the forecasters for the events in Lee et al.’s (2018) NFL dataset. Forecasters were ranked by performance in terms of percentage accuracy on each event. Error bars represent the standard error.

and meta-predictions in these datasets therefore appear to mostly reflect random noise rather than reflecting systematic differences in forecasters’ expertise.

Examining Figure 2.28, which shows the distribution of forecasters’ accuracy each week, we see that the plot for week 12 is somewhat bi-modal and therefore demonstrates evidence of heterogeneity in forecaster expertise for questions that week. There also appears to be a small subset of forecasters in week 17 who outperformed other forecasters in the crowd. Nonetheless, it’s not possible to draw any strong inferences about the distribution of forecaster expertise in individual weeks due to the limited sample size.

In general, these results are consistent with the results from our Experiment 1 and 2, which suggest that the SP mechanism is largely driven by systematic differences in expertise between sub-groups of forecasters in the crowd. It is interesting to see how poorly the SP algorithm performs when problems are genuinely predictive in nature. As the outcomes of NFL sporting matches are often hard to predict, this likely imposes a low ceiling on the best possible performance, unlike other datasets where ground truths exist. As a result, forecasters are likely to exhibit very

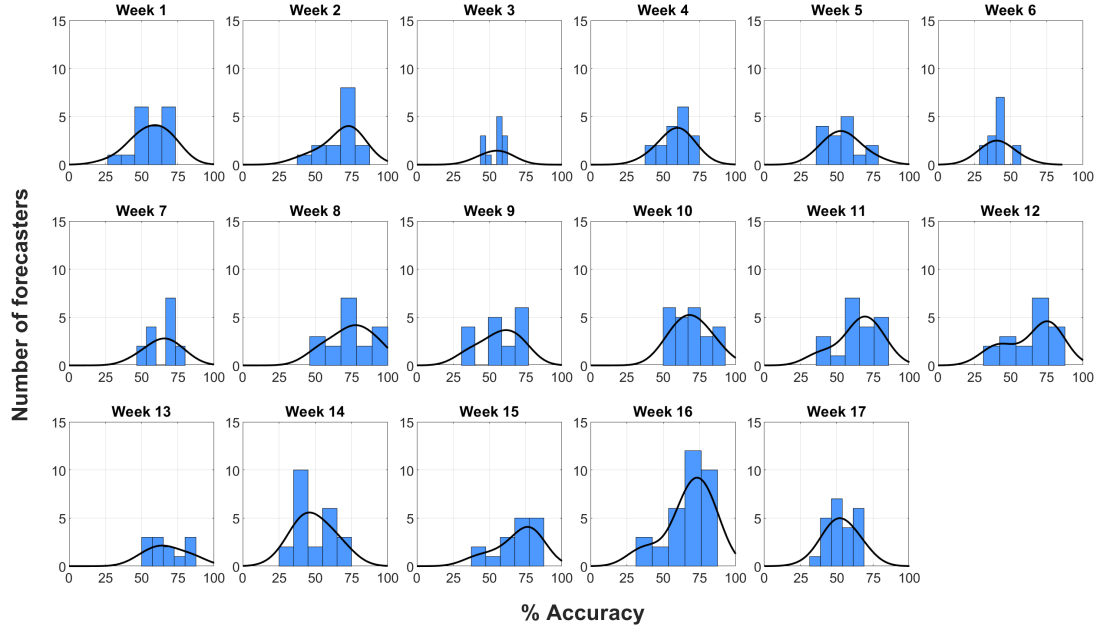


Figure 2.28: The distribution of forecasters over the percentage of events correctly predicted from the events in Lee et al.’s (2018)NFL dataset.

little systematic difference in expertise, and the SP algorithm is unlikely to provide substantial improvement over basic aggregation approaches such as majority voting. The efficacy of the SP algorithm thus appears to depend largely on the distribution of forecasters’ expertise in the crowd, and our results here support the general conclusions we have drawn thus far.

The results from these analyses, which demonstrate similar levels of performance for each algorithm, are useful in contributing to a wider collection of findings on the SP algorithm’s performance across different domains.

2.8 General Discussion

2.8.1 Contributions of the present research

The main contribution of this chapter is the identification and elucidation of a key mechanism of the Surprisingly Popular (SP) algorithm (Prelec et al., 2017). We identified a novel reformulation of the SP algorithm which highlights a fundamental mechanism that leverages experts’ predictions using

the absolute differences between forecasters’ votes and meta-predictions. Results from Experiment 1 and subsequent simulations provided preliminary evidence of this proposed mechanism, which we tested rigorously in Experiment 2. Results from Experiment 2 provided convergent evidence of this mechanism, providing insight into the environments under which the SP algorithm would be expected to outperform other forecast aggregation algorithms.

The current chapter explores the concept of expertise in two useful ways. Firstly, in our theoretical model, we define an individual’s expertise in the *ex ante* sense or in expectation – by the quality of information received by that individual, with experts receiving systematically more-informative signals than novices. The main proof for our theoretical model highlights that, under reasonable assumptions, the expected absolute differences between experts’ votes and meta-predictions will exceed that of novices. In this chapter, we show that the weights used by the SP algorithm also quantify expertise empirically, such that they are effective in distinguishing between high-performing individuals and low-performing individuals based on their actual responses.

We found that the weights used by the SP algorithm were effective measures of expertise under specific conditions. Our post-hoc analyses from Experiment 1 showed clear evidence that quantifying expertise using the SP weights was effective under conditions consistent with our theoretical model. Experiment 2 tested these predictions in a rigorous manner, and found that the SP weights quantified expertise most effectively under the conditions we proposed – questions of moderate difficulty, where crowd expertise was most heterogeneous and high in variability. The present chapter therefore provides evidence for a novel measure of expertise that has not been previously identified in the forecasting literature.

Our findings provide two other useful contributions: (1) Our theoretical model generates a wider range of predictions accounting for differences in empirical results from both Prelec et al.’s (2017) and our datasets. Our theoretical model is the first model in the literature to be able to account for these observed differences in the SP algorithm’s performance across domains. (2) Our replication using the same questions from Prelec et al.’s (2017) US states dataset demonstrates the robust replicability of the basic findings for the SP algorithm. Our re-analysis of Prelec et al.’s (2017) data also demonstrates the replicability of the results from their original analyses. Our results here constitute one of a small handful of cases in the literature where the SP algorithm

has been applied. Our demonstration for the replicability of these results therefore has important value in demonstrating the replicability and validity of Prelec et al.’s (2017) original findings.

2.8.2 Connection to existing research on the SP algorithm

The present chapter highlights a different mechanism underlying the SP algorithm than the one discussed in Prelec et al. (2017). Prelec et al.’s (2017) theoretical model showed the SP algorithm operated based on forecasters knowledge about the proportion of forecasters endorsing each outcome in factual and counterfactual worlds. However, this mechanism does not provide a clear account for the observed differences in the SP algorithm’s performance across different forecasting problems. To date, little guidance has been provided on the types of environments under which the SP algorithm should be applied, and the extent to which we can expect it to outperform alternatives in a given domain. The current chapter’s contributions in elucidating the SP mechanism therefore provides much-needed insight that was missing from the current literature.

Our proposed mechanism for the SP algorithm also accounts for more recent results published by Lee et al. (2018), who found that the SP algorithm offered no significant improvement over other standard forecast-aggregation algorithms for NFL prediction questions. Our analysis of their data at the dataset level showed that the crowd for each week was mostly homogeneous in expertise – an environment in which we have shown the SP algorithm to be typically ineffective. Our results can therefore account for the existing datasets with forecasters’ meta-predictions that are available in the current literature.

2.8.3 Connection to other expertise-identification approaches

The current chapter provides an important contribution in demonstrating the relationship between the SP algorithm to existing algorithms in the literature. While many past studies have sought to identify expertise by other measures, such from forecasters’ absolute or relative past performance (Budescu & Chen, 2015; Cooke, 1991; Mannes et al., 2014); forecasters’ variance, reliability, or inter-correlations (Davis-Stober et al., 2014; Genest & McConway, 1990); or consistency (Mellers, Baker, Chen, Mandel, & Tetlock, 2017); the current chapter demonstrates that forecasters’

expertise can also be identified by the absolute differences between forecasters' votes and meta-predictions. No study to date has identified the mechanism proposed in this chapter by which the SP algorithm identifies and leverages experts in the crowd. Our results thus provide valuable insight into how expertise can be reliably identified by using meta-cognitive knowledge when objective measures of expertise are unavailable and other subjective measures of expertise, such as confidence, are unreliable (Koriat, 2008, 2012).

The finding that the SP algorithm assigns weights corresponding to forecasters' latent expertise allows us to apply findings from other previous studies about the types of environments where other expertise-identification algorithms would be likely to outperform simple aggregation. For example, Mannes et al. (2014) found that aggregating over a small, select crowd outperforms a simple average over the whole crowd when there is high crowd dispersion (i.e., a large difference between the performance of the top experts versus the whole crowd). In the present work, crowd dispersion can be quantified (1) in theory, by the strength of garbling between experts' and novices' signals in our theoretical model (see Appendix - Section 6.1), and (2) empirically, by the distribution of forecasters' percentage accuracy for each dataset. The idea of heterogeneity of expertise discussed in the present paper therefore appears to be strongly connected to the concept of crowd dispersion discussed by these authors. Indeed, consistent with Mannes et al. (2014)'s findings, we see that the SP algorithm is similar to other expertise-based aggregation approaches in that it performs best when there are large differences in expertise between subsets of forecasters in the crowd. The findings on the SP mechanism in this paper therefore help to contextualise the research on the SP algorithm within the wider forecasting literature by demonstrating that the SP algorithm operates by identifying and extracting expertise much in the same way as many other contemporary forecast aggregation models (Budescu & Chen, 2015; Cooke, 1991; Mannes et al., 2014).

2.8.4 Considerations for future research

The current results suggest that the SP algorithm's improvement over other algorithms is limited to specific environments dependent on crowd expertise. While our results have demonstrated how knowledge of the SP algorithm's mechanism can be used to predict which environments it will

perform best, we were unable to find many examples of datasets where these improvements were as substantial as that for the US states grades dataset. Given the large range of datasets tested in the current chapter, including those from Prelec et al. (2017) and Lee et al. (2018), the range of problems where the SP algorithm would offer meaningful and consistent improvements over standard aggregation algorithms appears to be narrower than what is suggested by the current literature.

While this chapter has focused on categorical predictions for single-question forecasting problems, a natural way of extending the findings would be to incorporate forecasters' meta-predictions into a model that generates probabilistic predictions. Probabilistic models differ from categorical models in that they must quantify uncertainty in model predictions, thus making probabilistic models more attractive in applications of forecasting where the outcomes of events are often probabilistic in nature, such as sports betting and economic forecasting. While the SP algorithm generates only categorical predictions by design, its superior predictions in some domains such as the US states datasets may provide appealing advantages over other probabilistic forecasting models. The follow chapter thus explores the benefits and viability of adapting this innovative algorithm to probabilistic forecasting problems.

3 | Probabilistic single-question forecasting approaches

The work in this chapter was first written and then subsequently adapted, in part, for Martinie et al. (2020). The author-accepted version of the manuscript for this publication is included in the Appendix (see section 6.2).

The current chapter aims to extend previous work on categorical forecasting algorithms to probabilistic forecasting problems in the single-question domain, where the goal is to find the best aggregate prediction using forecasters' responses to only one question of interest. We examine the existing single-question probabilistic forecasting algorithms in the literature and provide a review of the wider range of theoretical and empirical issues underpinning the use of these approaches.

3.1 Probabilistic Forecasting

3.1.1 Quantifying uncertainty

In many forecasting applications, decision analysts want to know not only the most likely outcome for an event, but also the probability or likelihood of each outcome. For example, an automotive manufacturer might wish to forecast the probability that a particular component will malfunction in a particular time frame in their next batch of cars. The automotive manufacturer may need to make important business decisions such as the rate of production and pricing for different components, and so on, based on the probability of such malfunctions. A yes/no prediction of whether different components would malfunction may not be adequate for the automotive

manufacturer, since the business decisions about a component that malfunctions would also depend greatly on the certainty that those malfunctions would occur. Consider an extreme case comparing two different mechanical components: one has a 49% probability of malfunctioning in a particular time frame while the other component has a 1% probability of malfunctioning in the same period. A simple binary classification of whether each component is more or less likely to fail in that time frame would classify both components as less likely to malfunction than not. A binary forecast would thus be unable to capture a critical aspect of this forecasting problem, namely, the certainty associated with each outcome. In contrast, a forecast of the probability that each component will malfunction will be able to capture the differences in certainty between the two problems. Probability forecasts thus provide much greater discriminability between otherwise seemingly identical categorical predictions, due to the fact that the probability scale is continuous and discrimination between two different probability forecasts is only limited by the precision of measurement.

One approach for quantifying the certainty of forecasts (or confidence) is by the absolute difference between the forecasted probability and the baseline probability of 50%. Other scales, such as the log-odds scale, has also been shown to a good metric for quantifying uncertainty (Satopää et al., 2014). Regardless of scale, a key aspect of these metrics is that a forecast of an entirely uncertain event is predicted to occur with 50% probability, whereas an event whose outcome is entirely certain will be forecasted at a probability of 0% or 100%.

The log-odds scale has also been used to quantify uncertainty

This chapter will concentrate exclusively on probabilistic predictions about binary events, where there are two possible outcomes, typically a true-or-false prediction. In practice, multi-alternative problems can often be collapsed into two-alternative problems, for example, by collapsing multiple potential outcomes together into the same category. The algorithms discussed in this thesis are therefore potentially applicable in the multi-alternative problem domain as well, by first simplifying the decision problem.

3.1.2 The goal of probabilistic forecasting

The goal of probabilistic forecasting is to maximise *sharpness* subject to *calibration* (Murphy & Winkler, 1987). Calibration refers to the extent to which the observed frequency of outcomes agrees with the forecasted probabilities. For example, if someone forecasts that their favourite sports team will win with 50% probability for every match they played this year, and their team actually wins 50% of those games, they forecaster is said to be well-calibrated. From this example, it is clear that calibration is not a sufficient measure of performance, since that forecaster's performance might not accurately reflect their ability to differentiate between matches their team is likely to win or lose, only that they are aware of the overall probability of such outcomes. Probability forecasts also need to be sharp, such that they are as far away from baseline (50%) as possible while still being calibrated. A maximally-sharp forecaster would predict a probability of 0% or 100% for each event. Nonetheless, sharpness is not sufficient as the sole criterion for evaluating forecasts either, since a maximally-sharp forecaster may be randomly forecasting probabilities of 0% or 100% for each event in a way that does not correspond to the correct outcome, and will thus also be gravely miscalibrated. Calibration and sharpness are therefore both important criteria in determining the performance of probability forecasts. Brier scores, also known as the quadratic scoring rule, provide a combined measure of calibration and sharpness (Brier, 1950) and will be used extensively throughout this chapter and is discussed in the next section.

3.1.3 Scoring rules

Strictly proper scoring rules are conventional measures of performance in probabilistic forecasting and are widely used for both individual forecasts and aggregated or algorithm-based forecasts. Strictly proper scoring rules are useful because they ensure that performance of the probability forecasts, measured as some sort of score, is optimised only by forecasts of the true probability (Gneiting & Raftery, 2007). The use of scoring rules in assessing forecasts thus encourages forecasters to be careful and truthful in making their forecasts in order to maximise their score. Since the basic formulation of any scoring rule is sufficient for evaluating and comparing the performance of different probability forecasts, we will not discuss the decomposition of scoring rules

in further detail (for details on how scoring rules can be decomposed into separate components of reliability, resolution, and uncertainty, see Bröcker, 2009).

An infinite number of strictly proper scoring rules are possible (Murphy & Winkler, 1970) because any linear transformation of a strictly proper scoring rule always produces another strictly proper scoring rule. While the Brier and logarithmic scoring rules are both commonly used throughout the forecasting literature (e.g., Budescu & Chen, 2015; Mellers et al., 2017, 2015; Palley & Soll, 2019), the Brier scoring rule is typically preferred over the logarithmic scoring rule since the latter produces undefined scores for any forecasts which are maximally incorrect (i.e., forecasting a probability of 0% about the true outcome results in a score of $\log(0)$ and is therefore undefined). Perhaps for this reason, the Brier Score appears to be more commonly used. A linear transformation can be applied to the Brier score for ease of interpretability (e.g., Budescu & Chen, 2015), producing the transformed Brier scoring rule, which we use as the main measure of forecast performance throughout the next two chapters:

$$S = 100 - 50 \sum_{k=1}^K \frac{(D(o_k) - T(X_k))^2}{K},$$

where $D(o_k) = 1$ if the event is true and zero otherwise, and $T(X_k)$ is the probability assigned to that outcome being true by some algorithm or forecaster. This linear transformation of the Brier score retains the same functional form as the original, and is strictly proper (Murphy & Winkler, 1970). It also has a straightforward interpretation where scores range from 0 to 100, with 100 being a perfect forecast over all events and maximally uninformed forecasts of $p = .5$ receive a score of 75. Conveniently, this transformed Brier score is equivalent to a measure of percentage accuracy for events where forecasts are only either 0 or 1 (i.e., when applied to binary forecasts).

3.1.4 The Wisdom of Crowds literature

In this section, We provide a review of the forecast aggregation literature in the probabilistic domain.

The idea that combining information from multiple people can produce better predictions or decision outcomes has been well-known for centuries. Galton (1907) provided one of the first rigorous demonstrations of crowd wisdom in quantity estimation, in which he compared the error

associated with individual and crowd estimates of the weight of a butchered and ‘dressed’ ox at a livestock exhibition. Galton found that while forecasters were individually biased, the median forecast of the crowd was extremely accurate.

Since Galton (1907), the properties underlying crowd wisdom have become much better understood. The unweighted mean or simple average is the most common aggregation rule and has been shown to necessarily outperform the average forecaster when forecasts *bracket* the true probability (i.e., the range of forecasts fall above and below the true value; Larrick & Soll, 2006; Soll & Larrick, 2009). Furthermore, in the case that forecasts do not bracket the true value, the unweighted mean will perform just as well as the average individual’s forecast (Larrick & Soll, 2006). For example, if two people forecasted probabilities of 40% and 60%, and the true outcome occurred with a probability of 70%, then the average performance of those two forecasters will be necessarily equal or worse than the performance of the unweighted mean forecast for any strictly proper scoring rule. The unweighted mean is therefore theoretically always more effective than the performance of the average individual, and often, a substantial proportion of the crowd (Davis-Stober et al., 2014).

Under certain theoretical assumptions, the unweighted mean is the optimal aggregation algorithm for combining any set of forecasts. A well-known property of the arithmetic mean is that it is the maximum likelihood estimator of any population parameter under the assumption that samples are (1) drawn from a single distribution with one set of parameters (i.e., homogeneity), (2) any deviations from the mean population value is entirely random (i.e., errors are identically and independently distributed, or i.i.d), and (3) the shape of the distribution is symmetric and uni-modal. The maximum likelihood estimation property of the unweighted mean is important because it is guaranteed to produce the best estimate under these assumptions. For example, consider a set of observations drawn from a Gaussian distribution with unknown mean μ and variance σ^2 . If one was to estimate the μ of the Gaussian distribution that is most likely to have generated that set of observations, we would find that the estimated μ would always equal the unweighted mean of that set of observations. In forecast-aggregation terms, the unweighted mean will therefore necessarily outperform any other forecast, whether aggregated or from an individual, when all forecasters are drawn from the same symmetric, uni-modal distribution of expertise, and any differences in their prediction performance is due solely to random error. Note

that these conditions are sufficient for the unweighted mean to produce the optimal forecast, but not always necessary.

A similar aggregation model can be obtained by assuming the generating distribution to be binomial, such that the samples of forecasts from the distribution take values of either 0 or 1. Under the same assumptions of homogeneity and i.i.d errors, the maximum likelihood estimate of the probability parameter of the binomial distribution is given by the mean of those forecasts, or alternatively, the proportion of sample forecasts of 1. A binary version of this algorithm, which takes the estimated probability parameter and converts it into a binary vote, was discussed and applied in the previous chapter as the majority voting algorithm. Unsurprisingly, these models are rarely seen to outperform other, more sophisticated aggregation models (e.g., Budescu & Chen, 2015; Prelec et al., 2017), which have been developed to better model forecaster heterogeneity and correlated errors between forecasters.

Recent work by Davis-Stober et al. (2014) provides a formal definition of crowd wisdom that captures the exact conditions under which crowd wisdom will be observed. The authors develop a model that quantifies the trade-off between expertise, captured by forecasters' performance on questions with known outcomes, and diversity, captured by the correlation between forecasters' predictions. Their model demonstrates the way in which including non-expert forecasters can improve the aggregate prediction if those forecasters bring greater diversity to the crowd. Furthermore, the authors demonstrate that crowds are not necessarily wisest by simply aggregating the top performing individuals in the crowd, but rather, by optimising both the diversity of information that comprises the weighted sub-crowd and the sub-crowd's expertise. Unfortunately, their model requires that forecasters provide complete responses to every question in the dataset as well as forecasters' responses to multiple questions with known outcomes. As a result, this model may be inapplicable in many forecasting applications, since datasets often have missing data or records of forecasters' past performance on questions with known outcomes are unavailable. For this reason, their model is not directly applicable in the single-question domain.

3.1.5 Weighted forecast-aggregation models

Advanced methods for combining forecasts have since been developed that outperform basic aggregation approaches in environments where expertise is neither homogeneous nor i.i.d. These models rely on the assumption that forecasters' performance on questions of interest (i.e., *test events*) can be predicted reliably using forecasters' performance on a set of questions for which the outcomes are known and forecasters' performance can be quantified (i.e., *training events*), or some other variable, such as each forecaster's level of confidence. Clemen (1989), Clemen and Winkler (1999), and Genest and McConway (1990) provide well-summarised reviews of research developed up until the early 90s on different weighting approaches. These papers provide a comprehensive review of contributions from multiple fields including psychology, forecasting, statistics, and the social science literature, and discuss both theoretical and applied approaches to combining forecasts.

A popular approach that weights forecasters by their past performance is Cooke's (1991) 'classical' model, which borrows ideas from statistical hypothesis testing. In Cooke's model, forecasters are assigned a weight according to their calibration performance on a set of seed questions, and forecasters whose performance on the seed questions are below some given threshold, calculated using a null-hypothesis significance test, are assigned weights of zero, effectively removing their prediction from the crowd.

Later work by Mannes et al. (2014) compared the performance of different-sized groups of forecasters and found that selecting a small number of top-performing forecasters outperformed the unweighted mean prediction of the whole crowd across a wide range of settings. In an analysis of over 90 archival datasets, Mannes et al. (2014) showed that a select crowd of the five most knowledgeable judges yielded accurate and reliable judgments across a wide range of possible settings. Similar to Cooke (1991), their findings demonstrated how forecasters' performance on questions with known outcomes can be used as a measure of expertise, and thus used to improve forecasts on questions of interest.

More recently, Budescu and Chen (2015) developed the Contribution-Weighted Model, which selects and weights forecasters by their relative performance to other forecasters, rather than by forecasters' absolute performance. Forecasters were weighted by how much they improved the

(linearly transformed) Brier score of the aggregated forecast, rather than by their individual Brier scores. Conceptually, it may be easy to see why a measure such as “percentage of questions correct” is sometimes a poorer measure of expertise than “improvement in score over other forecasters”, since the former is contingent on which questions forecasters answered. The Contribution-Weighted Model therefore appears to be particularly effective in datasets where forecasters answered different subsets of questions, since experts provided the most improvement on questions where average performance was poor, but received similar Brier scores to non-experts who answered easier questions and therefore achieved the same performance. For example, imagine a forecaster who only makes predictions for questions where the majority of forecasters are correct, and another forecaster, identical in all other respects to the first forecaster but only makes predictions for questions where the majority of forecasters are incorrect. Both forecasters will receive the same score under the Brier score (and any other measure of raw performance, such as percentage accuracy) despite answering two different subsets of questions. In contrast, the Contribution-Weighted Model is effective at identifying expert forecasters who vote against the majority when the majority is wrong, since these forecasters provide the greatest contribution to the aggregate forecast.

3.1.6 Recalibration and extremisation approaches to forecast aggregation

While weighted aggregation approaches are effective at identifying expert forecasters in the crowd, they can often be improved by recalibration approaches that adjust for the overlap in shared information between forecasters (Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Turner, Steyvers, Merkle, Budescu, & Wallsten, 2014). Recalibration functions, or more typically, extremisation functions – adjust forecasts to more extreme values and are used to reduce the discrepancy between the forecasted probability and the observed frequency of occurrence (i.e., miscalibration). Recalibration functions can be used to transform both individual forecasts and aggregated forecasts, such as the unweighted mean, or more sophisticated weighted aggregation approaches. Recalibration functions are commonly used in data-driven applications of forecasting; popular choices include the logit and probit transformations (Baron et al., 2014; Satopää et al., 2014;

Satopää, Pemantle, & Ungar, 2016; Turner et al., 2014). Turner et al. (2014) provide an in-depth and nuanced comparison of different recalibration methods. Recalibration functions are typically symmetric around .5, such that probabilities are transformed solely to alter the certainty in the forecast, and are therefore generally ineffective when forecasts are categorically incorrect, however, more flexible recalibration approaches are able to bypass this constraint (Turner et al., 2014).

The need for extremisation – specifically, the finding that people generally over-estimate very small probabilities and under-estimate very large ones – has been demonstrated across a large number of the studies in the literature (for examples, see Baron et al., 2014; Dana, Atanasov, Tetlock, & Mellers, 2019; Shlomi & Wallsten, 2010; Turner et al., 2014; Zhang & Maloney, 2012). Several explanations have been proposed to account for the observed need for extremisation. Baron et al. (2014) provide a detailed exploration of two major contributing factors. The first factor relates to the compression of errors at the edges of the probability scale, due to the fact that the probability scale is bounded at 0 and 1. For events with outcomes that have high certainty (i.e., the actual probability of the outcome being true is close to 1 or 0), random error in how forecasters perceive and report their forecasts will result in an asymmetrical distribution in the reported forecasts. Even if random error is symmetrically distributed around the true probability, reported probabilities will always be bounded at the ends of the scale and thus the aggregate forecast will be regressed towards .5. This results in the mean of the observed forecasts always being biased relative to the true probability. The second factor relates to the overlap in shared information between forecasters. Suppose two forecasters each predict an event to occur with .8 probability. If each forecaster has access to different sets of information, then the aggregated probability forecast should in theory always be greater than .8, since the total amount of information across the two forecasters would exceed the amount of information that justified a .8 forecast. An unweighted mean of these two forecasts is therefore always going to be miscalibrated when forecasters have access to different sets of information.

More generally, any linear weighting of such forecasts cannot account for this overlap in information, and thus weighted aggregation approaches are also always miscalibrated when forecasters have access to different sets of information. Indeed, it has been shown that any weighted linear combination of probability forecasts is always theoretically miscalibrated and lacks sharpness (Ranjan & Gneiting, 2010). Thus, any aggregation algorithm that weights probability

forecasts, in theory, needs to be extremised in order to be correctly calibrated. Empirical findings seem to be consistent with the idea that extremisation improves the aggregated forecast for most aggregation approaches, and the practice of applying these recalibration functions to more-sophisticated aggregation approaches appears commonplace in the forecasting literature (e.g., see Casati, Ross, & Stephenson, 2004; Dana et al., 2019; Turner et al., 2014).

Although recalibration functions are effective at correcting for miscalibration, the standard approach to fitting recalibration functions cannot be directly applied to single-question forecasting problems. The parameters in recalibration functions, and the choice of function itself, are typically estimated on a *training set* of questions with known outcomes. The estimated function and parameters are then applied to a set of *test questions* – questions whose outcomes are unknown for the purposes of choosing the aggregation function or estimating the parameters of that function. It would therefore be impossible to apply recalibration functions that require this out-of-sample, parameter estimation approach to single-question forecasting problems.

Similarly, the cognitive modeling approach developed by Lee and Danileiko (2014) in theory also recalibrates forecasters based on a set of training data, however, without requiring the outcomes to those training to be known to the decision maker. This approach therefore has some advantage over other recalibration approaches, which require the outcomes to training questions to be known. Nonetheless, this information is also typically unavailable in the case of single-question forecasting problems, and thus the cognitive modeling approach cannot be directly applied to these problems.

A possible approach for single-question forecasting problems is to constrain the parameters in the recalibration function to be estimated from a different, unrelated dataset. Such an approach would allow decision analysts to recalibrate forecasts when the outcome is unknown and no other information regarding the forecasting problem is available. Importantly, this can be distinguished from other approaches that seek to identify the expertise of individual forecasters or specific properties of the forecasting problem, since a different set of forecasters are responding in each case and there may be no relationship between the questions of interest and the questions from which the parameters are estimated. Baron et al. (2014) provide a simple recalibration function with a single parameter which can be assumed to take on a fixed value:

$$t(p_k) = \frac{p_k^a}{p_k^a + (1 - p_k)^a} \quad (3.1)$$

where p_k is the original aggregated probability forecast for the k th event, $t(p)$ is the recalibrated probability, and a is the recalibration parameter, which determines the strength of the transformation. This function, which has been used by Baron et al. (2014); Erev, Wallsten, and Budescu (1994); Shlomi and Wallsten (2010); Turner et al. (2014); and others before them, extremises probability forecasts when $a > 1$ and anti-extremises when $0 < a < 1$. Baron et al. (2014) found that this function worked well in crowds containing experts at approximately $a = 2.5$, and in crowds containing non-experts, approximately $a = 3.5$. Since recalibration approaches have not been applied in the single-question domain previously, we adopted the more conservative parameter value, 2.5, to recalibrate the forecasts from different single-question algorithms. One of the aims of this chapter is therefore to test the effectiveness of the recalibration approach in the single-question domain, particularly when combined with existing single-question aggregation approaches.

3.1.7 Single-question aggregation approaches

Another limitation of weighted aggregation models is that they typically require forecasters' past performance, or some function of their past performance, to select and weight subsets of the crowd. In many forecasting problems, forecasters may only provide responses to a single question (e.g., due to time or monetary constraints), and so it would therefore be impossible to identify expertise using these models since this information is unavailable. In other cases, the forecasting problem might be so novel that records of forecasters' performance on other questions might not effectively predict their performance on questions of interest. Nonetheless, in those scenarios, decision makers may still look to be able to identify and leverage the expertise of forecasters in the crowd. Measures like self-reported expertise or confidence are often poorly correlated with actual expertise (Cooke, 1991; Koriatic, 2008, 2012; Kruger & Dunning, 1999; Prelec et al., 2017), thus there is a clear need for other methods of identifying expertise.

Several approaches have been developed for aggregating probabilistic forecasts in single-question forecasting problems in recent years. Two notable approaches include the P''_{cs} aggregator,

as part of the Gaussian Partial Information model proposed by Satopää et al. (2016), and the Minimal Pivoting approach proposed by Palley and Soll (2019). Like the model proposed by Davis-Stober et al. (2014), both these algorithms take a theoretically-driven approach to model the overlap in information, however, the P''_{cs} aggregator and Minimal Pivoting model make stronger theoretical assumptions about the structure of information received by forecasters, allowing these algorithms to generate probability forecasts without needing records of forecasters' past performance on training questions.

Satopää et al.'s (2016) Gaussian Partial Information Model

In their paper, Satopää et al. (2016) develop an algorithm that *extremises* probability forecasts, such that forecasts are shifted closer to their nearest end of the probability scale. For example, extremisation of a forecasted probability of $p = .9$ involves a transformation of that probability: $t(p) \in (.9, 1]$. The algorithm Satopää et al. (2016) propose makes use of the fact that forecasters often have overlapping sources of information when generating forecasts. Standard aggregation approaches such as the unweighted mean are unable to account for this overlap in information, since they often assume all forecasters have access to the same set of information, argued by the authors to be both implausible and unlikely to be true in practice. In particular, the unweighted mean has no mechanism to account for the diversity of information in the crowd, and is therefore generally an unsuitable aggregator in practice. In response, the authors propose the *Gaussian Partial Information Model*, where units of information are represented as Gaussian particles providing evidence in favour of or against some outcome. While the parameters of the optimal aggregator under this Gaussian model can only be estimated using forecasters' responses on multiple questions, a restricted version of the model can be used to generate forecasts for single-question forecasting problems. By assuming that forecasters' information structure is compound symmetric – where forecasters have the same amount of information and each possible pair of forecasters have the same overlap in information – the parameters of the optimal aggregator can be estimated from forecasters' probability forecasts on just one question. The authors demonstrated that their proposed P''_{cs} aggregator outperforms simple averaging on a real-word forecasting dataset from the forecasting ACE tournament (see Satopää et al., 2016, for details). Their model thus represents an effective theoretically-driven approach to generating probability forecasts in the

single-question domain. In the current chapter, we will examine how the P''_{cs} aggregator performs relative to other contemporary forecast-aggregation approaches in the single-question domain.

Palley and Soll's (2019) Minimal Pivoting model

Another approach to the shared information problem between forecasters is the Minimal Pivoting model proposed by Palley and Soll (2019). In their paper, the authors develop a theoretical framework where they model the structure of information available to forecasters in the crowd. Similar to Satopää et al. (2016), their framework draws on the idea that the unweighted mean will perform sub-optimally when the information available to forecasters is shared. Another way to conceptualise this problem is that when multiple forecasters base their forecasts on the same information, that information will become over-represented in the crowd aggregate. In contrast, unique information available to only a small subset of forecasters will become under-represented. Using this framework, the authors develop a series of algorithms that correct for this bias from information.

The optimal algorithm for combining forecasts when information is shared depends heavily on the structure of the information shared by forecasters in the crowd. The authors proposed three general, idealised types of information structures that describe how information can be shared between forecasters: (1) symmetric, in the case where all forecasters share some common information but also have access to their own private information; (2) nested-symmetric, which is similar to the first case except some forecasters (i.e., laypeople) only have access to the shared information, and no private information; (3) and nested, where the crowd is comprised of laypeople as well as experts – people who have access to private information (i.e., information that is not available to laypeople) but this information is shared amongst all the experts. Since the exact information structure available to forecasters might be unknown to the decision analyst seeking to aggregate these forecasts, the authors proposed a reasonable heuristic to be a Minimal Pivoting approach, which provides the most conservative correction relative to the three other pivoting procedures discussed in their paper.

In order for the shared information to be identified in the pivoting model, forecasters must first provide meta-predictions about the average forecast of other forecasters. The Minimal Pivoting model works by using forecasters' meta-predictions to identify the information that is shared

between forecasters and the information that is private to a subset of forecasters, which the model then combines into one optimal, aggregated forecast. The Minimal Pivoting model is formalised as:

$$T_{MP}(X_k) = \sum_{i=1}^{N_k} \frac{P_{i,k} + (P_{i,k} - M^P_{i,k})}{N_k}$$

where $T_{MP}(X_k) \in [0, 1]$ is the Minimal Pivoting model’s forecast for the k th event, $P_{i,k}$ is the probabilistic prediction of the i th forecaster for the k th event, N_k is the number of forecasters for the k th event, and $M^P_{i,k}$ is the i th forecaster’s prediction about the average probability forecast across all other forecasters for the k th event.

Palley and Soll (2019) tested the Minimal Pivoting model across a range of simulated and empirical settings over four experiments, including a dataset of forecasts for the outcomes of NCAA basketball games. The Minimal Pivoting model generally outperformed the unweighted mean in terms of forecasting accuracy, relative error, and mean Brier score. The Minimal Pivoting model was therefore an effective alternative to the unweighted mean for aggregating probabilistic forecasts in the single-question domain.

The Minimal Pivoting model is similar to the Surprisingly Popular (SP) algorithm proposed by Prelec et al. (2017) in the sense that both these algorithms use forecasters’ meta-cognitive knowledge to aggregate forecasts. However, there are several key differences between the two algorithms. At a basic level, the two algorithms differ in the type of meta-prediction response they use. The SP algorithm uses forecasters’ meta-predictions about the proportion of other forecasters voting “true”, M^V , whereas the Minimal Pivoting model uses forecasters’ estimates of the average probability forecast predicted by others, M^P . The theoretical basis underlying each algorithm also differs fundamentally. The Minimal Pivoting model was developed to account for shared information according to the different types of information systems available to forecasters. In contrast, as we demonstrated in Chapter 2, the SP algorithm operates by identifying and leveraging forecasters’ expertise, which can be identified from the absolute differences between forecasters’ votes and meta-predictions. In principle, the SP algorithm is thus highly similar to other weighting algorithms in the forecasting literature, while the Minimal Pivoting model does not assign weights to forecasters at all.

Additionally, the two algorithms also differ in that the SP algorithm has a bias-correcting property, such that it can override outcomes endorsed by a large majority of the crowd (i.e., the SP algorithm can make any possible prediction regardless of the strength of the aggregated forecast). In contrast, the Minimal Pivoting model is only able to anti-extremise (i.e., shift the forecast closer to the baseline probability of .5) when the unweighted mean forecast is close to the ends of the probability scale. Specifically, when the unweighted mean of all forecasts, $\sum_{i=1}^{N_k} P_{i,k}$, is greater than .75 (or less than .25), then pivoting is no longer able to alter the forecasted outcome from true to false (and vice-versa), since the average meta-prediction cannot exceed 1. In contrast, the SP algorithm is able to switch from the outcome endorsed by the majority regardless of the percentage of forecasters endorsing that outcome, since the average meta-prediction can always be greater than or equal to the proportion of forecasters voting true in any problem where crowd vote is not completely unanimous. The Minimal Pivoting is therefore constrained on how far its prediction can deviate from the average forecast. In the current chapter, we will compare the efficacy of the Minimal Pivoting model with the SP algorithm and other single-question aggregation approaches in the probabilistic domain.

Aims of the present research

The current chapter has two primary research aims:

- Currently, there remains no effective way for identifying and leveraging the expertise of forecasters in the single-question domain. Although the P''_{cs} aggregator and Minimal Pivoting model both generate accurate forecasts without the need for records of forecasters' past performance, they do not allow decision makers to quantify forecasters' expertise. One of the central aims of this chapter is therefore to develop a measure by which forecasters' expertise can be robustly quantified and thus leveraged to produce accurate aggregated forecasts.
- While the P''_{cs} aggregator and Minimal Pivoting model appear to perform well empirically relative to the other basic aggregation approaches such as simple averaging, it remains unclear whether these two aggregation approaches will outperform each other, and whether they will outperform other single-question aggregation approaches such as the SP algorithm.

There have been no comparisons made between the P''_{cs} aggregator, the Minimal Pivoting model, or the SP algorithm in the literature to date. The current chapter aims to provide an empirical comparison between these models for aggregation in the single-question domain and present potential avenues for extending the findings from Chapter 2 to the probabilistic domain.

The existing datasets in the literature do not allow for a fair comparison of the Minimal Pivoting model and SP algorithm. No studies to date have elicited both forecasters' meta-predictions about the average prediction of others, which the Minimal Pivoting model requires, as well as forecasters' meta-predictions about the proportion of others voting "true" for each outcome, which the SP algorithm requires. Furthermore, in order to accurately gauge how well each algorithm will perform generally when the forecasting domain is unknown, a sufficiently-large and diverse set of problems is needed.

One solution might be to assume that forecasters' responses to both questions are identical, which will allow for comparisons between these algorithms on datasets that are currently available in the literature (e.g., those from Chapter 2). While there are important theoretical differences between these two types of responses, we found in pilot testing the differences forecasters' responses to these two questions to be generally very small. For the purposes of a preliminary comparison of both algorithms on existing datasets in the literature, we therefore assume that forecasters' responses to both these meta-prediction questions in practice to be identical. Given that these datasets from the previous chapter comprise only questions where forecasters provide meta-predictions about the proportion of forecasters voting "true", rather than meta-predictions about the average probability forecasted by others, this is likely to disadvantage the Minimal Pivoting model, since it was not developed to use these responses. Thus, while we make the comparison between the Minimal Pivoting model and other probabilistic algorithms using these responses, we note that the difference in performance is likely to reflect a lower bound on the performance of the Minimal Pivoting model. In Experiment 4 of this chapter, we conduct an experiment which provides a comprehensive comparison of Minimal Pivoting model to other aggregation approaches by using the appropriate responses for each algorithm. Indeed, in that experiment, we find that (1) our general findings using the datasets from Chapter 2 are reliably replicated, and (2) empirically,

based on the similar pattern of results observed in Experiment 4 compared to Experiments 1-3, that forecasters’ meta-predictions about the proportion of forecasters voting “true’ provide a fairly robust approximation of their meta-predictions about the average probability forecast of others.

3.2 Experiment 1: Testing categorical forecast-aggregation approaches in the probabilistic forecasting domain

Since the recent development of the P''_{cs} aggregator and Minimal Pivoting model, these aggregation approaches have yet to be compared empirically. It remains unclear whether one of these algorithms offers better probabilistic forecasts in practice. In this section, we therefore provide an empirical comparison of these two approaches using the datasets studied in Chapter 2. We will also examine the performance of categorical algorithms from Chapter 2, by treating the binary, aggregated forecasts from each algorithm as probability forecasts. As no studies have reported applying these categorical algorithms to probabilistic forecasting problems, it remains unknown how these algorithms will perform. The categorical forecasting algorithms discussed in the previous chapter are equivalent to probabilistic forecasting algorithms where the aggregated forecast has been maximally extremised, such that only probabilities of 1 or 0 are generated by each algorithm. As such, they may perform well on probabilistic forecasting problems where forecasters are systematically under-confident, as suggested by the observed need for extremisation seen in the literature (e.g., Baron et al., 2014; Satopää et al., 2016; Turner et al., 2014). In particular, it is unknown whether the SP algorithm, which outperformed other algorithm on several datasets under binary measures of performance, is able to perform well under probabilistic measures of performance.

Binary algorithms may perform poorly on probabilistic measures of performance for at least two reasons. Firstly, since these algorithms cannot produce non-binary forecasts, they are likely to be miscalibrated on events which are highly uncertain (i.e., the true outcome occurs with a probability close to .5). Secondly, binary algorithms such as these often have no mechanism to make use of the uncertainty in forecasters’ predictions. For example, the SP algorithm and majority voting algorithm do not account for forecasters’ confidence at all. As such, these

algorithms are likely to perform poorly on probabilistic forecasting problems because they do not account for uncertainty in forecasters’ predictions inherent in probabilistic forecasting problems. In contrast, two of the categorical forecasting algorithms discussed in the previous chapter – the Confidence-weighted algorithm, which weights forecasters’ votes by their confidence, and the Max confidence algorithm, which selects the outcome with the highest average confidence – can naturally account for forecasters’ uncertainty in the aggregate prediction. Nonetheless, these algorithms ultimately produce a binary forecast, which may be disadvantageous when outcomes occur with high uncertainty. Given the conflicting factors that may determine the performance of these categorical algorithms on probabilistic measures of performance, an empirical comparison between these algorithms would provide valuable insight into the issue.

3.2.1 Methods

We tested the performance of different aggregation approaches across a large collection of 1196 events from all 12 datasets reported in Chapter 2. The 12 datasets include: five datasets kindly provided by Prelec et al. (2017), six datasets we collected in Experiments 1 and 2 of Chapter 2, and a large dataset comprising 17 weeks of NFL forecasts kindly provided by Lee et al. (2018). The full details of the types of questions in each dataset can be seen in Table 3.1.¹

For each of the 1196 events, forecasters provided a vote about the most likely outcome, the probability that they believe that outcome would occur, and either the proportion of other forecasters predicting that “true” would be the correct outcome (for datasets 1-5, which were collected by Prelec et al., 2017, and for datasets 6-11, which we collected), or the proportion of other forecasters that would agree with the forecaster’s answer (for dataset 12, which was collected by Lee et al., 2018). The meta-predictions for the NFL were transformed to be commensurate to with meta-predictions in the other 11 dataset by inverting the responses where forecasters voted “false”. Thus, the structure of responses to all 1196 events were made identical. We collapsed all 17 datasets into one large “dataset”.

¹As an initial investigation analysis, we first analyse the aggregate-level performance of these aggregation approaches. As these datasets contain multiple different problem domains and different types of problems, it may be inappropriate to average over qualitatively different types of problems. For this reason, we also provide analyses at a dataset level in a later section of this chapter.

Table 3.1: Details for each dataset used in this chapter

Dataset (No. of Questions)	Description	Example	Source
US States - PSM (50)	The capital city of each US state	Birmingham is the capital of Alabama	Prelec et al. (2017)
Trivia (80)	General-knowledge trivia questions	The Empire State Building has its own zip code	Prelec et al. (2017)
Lesions (80)	Dermatologists diagnosing 80 skin lesion images as benign or malignant	This lesion is benign	Prelec et al. (2017)
Art - Novices (90)	Laypeople asked to predict the market price category of artworks	This artwork is worth more than \$30,000	Prelec et al. (2017)
Art - Experts (90)	Experts asked to predict the market price category of artworks	This artwork is worth more than \$30,000	Prelec et al. (2017)
US States - MWH (50)	Replication of Prelec et al.’s (2017) US states dataset	Birmingham is the capital of Alabama	Current thesis
US Grades 1 (100)	General science questions from grades 1 and 2	Rabbits are omnivores	Current thesis
US Grades 2 (100)	General science questions from grades 3, 4, and 5	A lack of Vitamin C causes scurvy	Current thesis
US Grades 3 (100)	General science questions from grades 6, 7, and 8	Ultraviolet light is invisible to the human eye	Current thesis
US Grades 4 (100)	General science questions from grades 9 and 10	Elements in the standard periodic table are arranged in terms of atomic mass	Current thesis
US Grades 5 (100)	General science questions from grades 11 and 12	In kinetic particle theory, all collisions within a system are assumed to be elastic	Current thesis
NFL predictions (256)	Predictions for the NFL games in the 2017-2018 season	Denver Broncos will win against the Dallas Cowboys this Sunday	Lee et al. (2018)

3.2.2 Individuals' responses

Adopting the same notation as the previous chapter, we consider a series of K events indexed by $k = \{1, \dots, K\}$ each with a binary outcome $o_k \in \{T, F\}$. We say an event is true if $o_k = T$. For each event k , a crowd of N_k forecasters is assembled. From each forecaster $i \in \{1, \dots, N_k\}$ we elicit three reports: the forecaster's prediction of whether an event is true, $V_{i,k} \in \{0, 1\}$, the forecaster's estimate of the probability that the event is true $P_{i,k} \in [0, 1]$, and the forecaster's meta-prediction about the votes of others, which is the forecaster's prediction of the proportion of the crowd who predict that the event is true, $M_{i,k}^V \in [0, 1]$. In the final experiment in this chapter, we also elicited the forecaster's meta-prediction about the average probability forecasted by others, $M_{i,k}^P \in [0, 1]$. We let $X_{i,k} := (V_{i,k}, P_{i,k}, M_{i,k}^V, M_{i,k}^P)$ be forecaster i 's full report and assume that the reports are consistent in the sense that $V_{i,k} = 0$ if $P_{i,k} < 0.5$ and $V_{i,k} = 1$ if $P_{i,k} > 0.5$. The way we elicited the predictions ensured that this assumption was never violated, except in the final experiment in this chapter, where we excluded any forecasts for which this violation occurred. We assume that if $P_{i,k} = 0.5$, forecasters vote randomly for either outcome with equal probability.

Let $X_k = \{X_{i,k}\}_{i=1}^{N_k}$ be the full set of reports for event k . Each algorithm we consider aggregates a set of individual forecasts into a single probability forecast $T : X_k \rightarrow [0, 1]$. As discussed in the introduction, we restrict attention in this chapter to single-question algorithms, which only use the reports for event k to form the forecast for that event.

3.2.3 Algorithms

We are primarily interested in how the SP algorithm will perform in the probabilistic domain, relative to the P''_{cs} aggregator (Satopää et al., 2016) and Minimal Pivoting model (Palley & Soll, 2019). For completeness, we also include each of the categorical forecasting algorithms from the previous chapter. As a benchmark, we compare the performance of each algorithm to the unweighted mean. Our full comparison set of algorithms includes the majority voting algorithm, the Confidence-weighting algorithm, the Max Confidence algorithm, the SP algorithm,

the unweighted mean, the recalibrated unweighted mean², the P''_{cs} aggregator, and the Minimal Pivoting model (for equations, see tables 3.2 and 3.3).

As inferential tests for statistically significant differences in performance between algorithms, we use the empirical bias-corrected and accelerated bootstrap (Efron, 1987; Efron & Tibshirani, 1994), which is analogous to the paired-samples t-test but does not make any parametric assumptions about the population. We compute 95% confidence intervals (CIs) for each comparison, which indicate whether the difference between their performance is statistically significant at the $\alpha = 5\%$ level when the interval values exclude the null hypothesis value (for all comparisons in this thesis, $H_0 = 0$).

²Note that as a preliminary test of the performance of these aggregation approaches, the only approach we recalibrated was the unweighted mean. This is because neither the P''_{cs} aggregator and Minimal Pivoting model are weighting approaches and therefore in theory do not need to be extremised. In the subsequent experiment, we test whether these other aggregation approaches could also benefit from extremisation.

Table 3.2: Binary aggregation approaches

Algorithm Name	Formula	Description
Majority Vote	$T_{MV}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{V_{i,k}}{N_k} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$	Select the answer with the most votes
Confidence-weighted	$T_{CW}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{P_{i,k}}{N_k} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$	Select the answer with the greatest confidence-weighted vote
Max confidence	$T_{MC}(X_k) = \begin{cases} 1 & \text{if } \sum_{\{i V_{i,k}=1\}} \frac{P_{i,k}}{N_{t,k}} > \sum_{\{i V_{i,k}=0\}} \frac{1 - P_{i,k}}{N_{f,k}} \\ 0 & \text{otherwise.} \end{cases}$	Select the answer with the greatest mean confidence (Note: $N_{t,k} = \sum_{i=1}^{N_k} V_{i,k}$; $N_{f,k} = N_k - N_{t,k}$)
Surprisingly Popular	$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{ V_{i,k} - M_{i,k} V_{i,k}}{\sum_{j=1}^{N_k} V_{j,k} - M_{j,k} } > 0.5 \\ 0 & \text{otherwise.} \end{cases}$	Select the answer that is surprisingly popular

Table 3.3: Probabilistic aggregation approaches

Algorithm Name	Formula	Description
Unweighted mean	$T_{UM}(X_k) = \sum_{i=1}^{N_k} \frac{P_{i,k}}{N_k}$	Unweighted average of all individual forecasts
Recalibrated unweighted mean	$T_{UM-R}(X_k) = \frac{T_{UM}(X_k)^{2.5}}{T_{UM}(X_k)^{2.5} + (1 - T_{UM}(X_k))^{2.5}}$	Recalibrated version of the unweighted mean (see Baron et al., 2014)
P''_{cs} aggregator	$T_{P''_{cs}}(X_k) = \Phi\left(\frac{\frac{1}{(N-1)\lambda+1} \sum_{i=1}^N X_{B_i}}{\sqrt{1 - \frac{N\delta}{(N-1)\lambda+1}}}\right)$	Revealed Aggregator for the Gaussian Model under compound symmetry. Where λ is the estimated amount of information used by each forecaster, δ is the estimated overlapping proportion of information, and $X_{B_i} = \Phi^{-1}(p_i)\sqrt{1 - \delta}$ for all $i = 1, \dots, N$. (see Satopää et al., 2016, for technical details).
Minimal Pivoting	$T_{MP}(X_k) = \sum_{i=1}^{N_k} \frac{P_{i,k} + (P_{i,k} - M^P_{i,k})}{N_k}$	Unweighted mean corrected by the Minimal Pivoting procedure (see Palley & Soll, 2019)

To provide deeper insight into the performance of these algorithms, we also generated calibration curves for each of the non-binary algorithms’ forecasts over all events. The calibration curve for an algorithm shows the forecasted proportion of events to be “true” compared to the actual proportion of events where the outcome was “true”. In our plots, the events forecasted at each bracket of probabilities (e.g., 0-10 % of being “true”) are collapsed into a single bin, and the y-value of that bin indicates the proportion of those events for which the outcome was actually “true”. The diagonal line thus indicates the calibration curve of a perfectly-calibrated forecaster (or algorithm) whose forecasted proportion of events to be “true” matches exactly the observed proportion of “true” events. Deviations from the diagonal line indicate miscalibration, with points falling above the diagonal line indicating under-confidence, and points falling below the line indicating over-confidence.

3.2.4 Results and Discussion

Figure 3.1 shows the mean transformed Brier score for each of the binary algorithms, compared to the unweighted mean, the Recalibrated unweighted mean, the P''_{cs} aggregator and the Minimal Pivoting model. Figure 3.2 shows the calibration curves for each of the probabilistic algorithms. We omit the calibration plots for each of the binary algorithms since they do not make any forecasts except in the first and last bins.

From Figure 3.1, we can see that altogether the binary algorithms perform poorly compared to the probabilistic algorithms in terms of mean score. Conveniently, the transformed Brier score used here is equivalent to percentage accuracy for binary forecasts. The Brier score for each of these algorithms is therefore a linear composite of each algorithm’s percentage accuracy on the datasets shown in figures 2.3, 2.16, and 2.25. The SP algorithm performed significantly worse than the unweighted mean by 5.52 points in Brier score (95% CI for bootstrap mean paired difference: [3.56, 7.43]). Similarly, the Majority Vote, Confidence-weighted, and Max confidence algorithms also performed significantly worse than the unweighted mean (95% CI: [6.56, 10.35], [5.59, 9.27], and [7.08, 11.31], respectively). The SP algorithm also performed significantly worse than the Recalibrated unweighted mean, the P''_{cs} aggregator, and Minimal Pivoting model (95% CI: [4.06 7.48], [3.57, 7.34], and [4.54 8.00], respectively). The standard SP algorithm appears

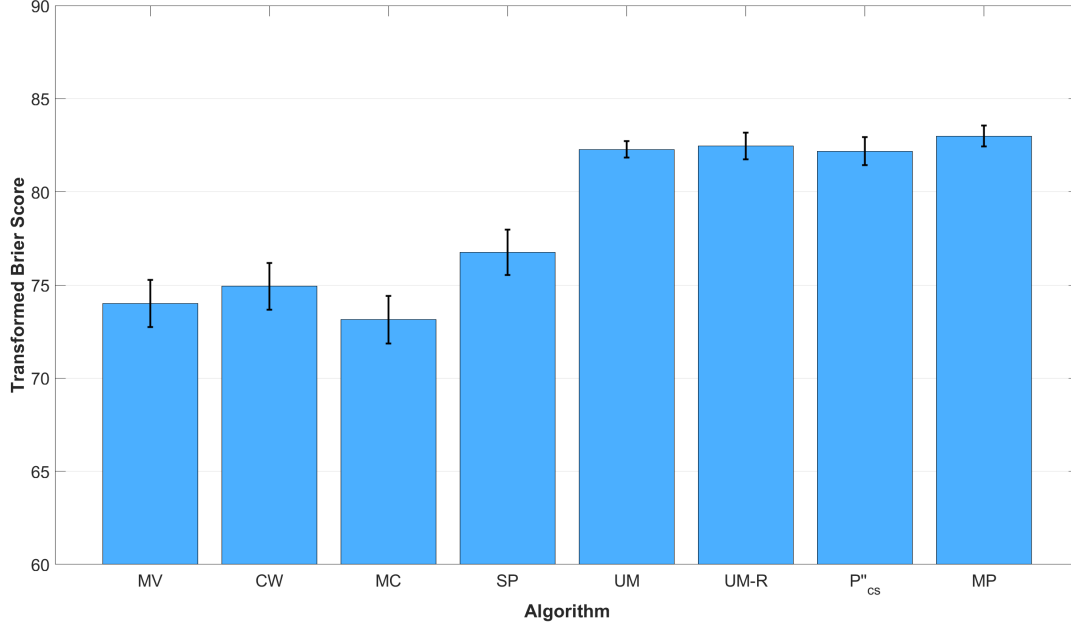


Figure 3.1: The mean transformed Brier score for the Majority Vote (MV), Confidence-weighted (CW), Max confidence (MC), and Surprisingly popular algorithm (SP), the unweighted mean (UM), Recalibrated unweighted mean (UM-R), the P''_{cs} aggregator (P''_{cs}), and the Minimal Pivoting (MP) algorithm, on 1196 events across 12 datasets. Error bars show the standard error.

to perform extremely poorly in the probabilistic domain, and cannot be applied effectively to probabilistic forecasting in its current form.

The difference in score between these binary algorithms and probabilistic algorithms is largely accounted for by the fact these algorithms only produce binary forecasts. For example, the only difference between the equation for the confidence-weighted algorithm and the unweighted mean is the final extremisation step in the confidence-weighted algorithm. The difference in Brier score between the confidence-weighted algorithm and the unweighted mean of 7.35 points (95% CI: [5.59, 9.27]) is thus entirely due to extremisation. While no equivalent probabilistic algorithms exist for these other binary algorithms, the fact that each of these binary algorithms performed worse than the probabilistic algorithms to a very similar extent suggests that these binary algorithms perform poorly due to similar reasons.

Interestingly, all of the probabilistic algorithms obtained a similar range of Brier scores. Extremisation of the unweighted mean using equation 3.1 (adapted from Baron et al., 2014)

resulted in a small and non-significant increase of 0.18 points in Brier score (95% CI: [-0.46, 0.79]). The reason the recalibration function was not effective at improving the aggregated forecasts could potentially be because the recalibration function was not appropriately optimised for this general set of questions. However, this highlights a potential limitation of trying to apply recalibration functions in the single-question domain, since it would not be possible to estimate the parameters in an out-of-sample manner for these forecasting problems.

The calibration curve for the unweighted mean, shown in Figure 3.2 as the light-orange line, appears close to the reference diagonal line and is therefore fairly well-calibrated. In contrast, the Recalibrated unweighted mean, shown as the red line, appears only slightly closer to the diagonal line in the most extreme bins either side (the 0 - 10, 10 - 20, 80 - 90, and 90 - 100 percentile bins), but almost as equally divergent as the raw unweighted mean in most of the other bins. Thus, while the unweighted mean was not particularly well-calibrated, the recalibration function was ineffective at correcting for miscalibration in most bins.

As Figure 3.1 shows, the P''_{cs} aggregator performed slightly (but not significantly) worse than the unweighted mean (95% CI: [-0.74, 0.98]). The calibration curve in Figure 3.2 shows that the P''_{cs} aggregator was indeed more miscalibrated than the unweighted mean, Recalibrated unweighted mean, and Minimal Pivoting model. The P''_{cs} aggregator displays the opposite pattern of miscalibration compared to the unweighted mean, where the aggregated forecast is over-confident, rather than under-confident, and therefore requires anti-extremisation to be calibrated. For example, the forecasts in the 0–10, 10–20, 20–30, 50–60, and 60–70 percentile bins all benefit from anti-extremisation. This is perhaps not surprising, given that the P''_{cs} aggregator is designed to produce an extremised forecast under the assumptions of compound symmetry. These results therefore suggest that forecasters’ information structure was not compound symmetric – forecasters were likely to have had access to different sources of information, such that they would differ in expertise, resulting in the algorithm over-extremising. These results are consistent with the findings from Chapter 2, which showed that many of these datasets contained groups of forecasters heterogeneous in expertise (e.g., see Figure 2.11), and therefore likely to have different information structures.

While the Minimal Pivoting model was not developed to use forecasters’ meta-predictions about the proportion of other forecasters voting “true”, the algorithm still outperformed all other

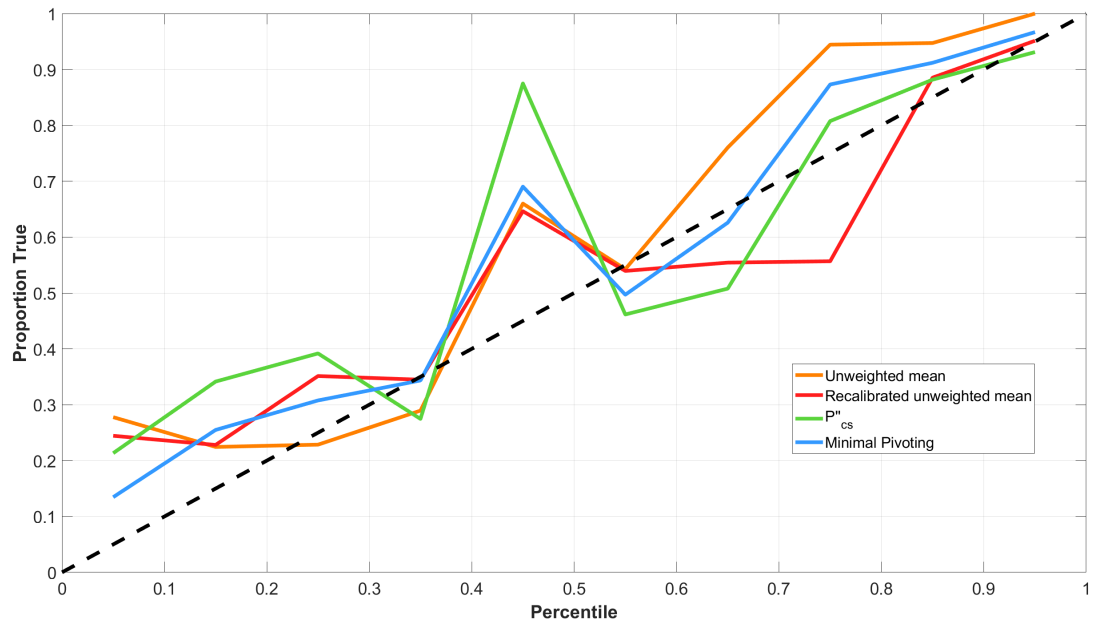


Figure 3.2: The calibration curve for the unweighted mean (orange), the Recalibrated unweighted mean (red), the P''_{cs} aggregator (light green), and the Minimal Pivoting model (blue). Each calibration curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the calibration of a perfectly-calibrated forecaster (or algorithm) whose forecasted probabilities match exactly the proportion of “true” events observed.

algorithms, as Figure 3.1 shows. The Minimal Pivoting model significantly outperformed the SP algorithm by 6.24 points (95% CI: [4.56, 8.07]) and the unweighted mean by 0.27 points (95% CI: [0.20, 1.23]). However, it did not significantly outperform the Recalibrated unweighted mean (95% CI: [-0.185, 1.286]) nor the P''_{cs} aggregator (95% CI: [-0.12, 1.76]). Indeed, Figure 3.2 shows that the Minimal Pivoting model was better calibrated overall compared to the other probabilistic algorithms. The Minimal Pivoting model therefore provided the most accurate probabilistic predictions amongst these algorithms.

The Minimal Pivoting model’s impressive performance on these datasets – despite using meta-prediction responses designed for the SP algorithm – suggests that it may have been able to perform even better if we had elicited forecasters’ meta-predictions about the average forecast of all other forecasters as well, since that is the type of response the model was designed to handle. Nonetheless, given that no studies to date have reported both types of meta-predictions, it remains unclear whether the Minimal Pivoting model’s improvement over other probabilistic-forecasting algorithms is likely to be much larger than the difference observed in this large dataset. More concretely, although the Minimal Pivoting model outperformed other aggregation approaches, the improvement it offered over the unweighted mean was very small (approximately .27 points in transformed Brier score). Other probabilistic forecasting approaches such as extremisation and the P''_{cs} aggregator offered little to no improvement over the unweighted mean. Overall, these results suggest that more effective aggregation approaches are needed for generating better predictions to single-question forecasting problems.

One potential solution could be to adapt the weighting function used by the SP algorithm to the probabilistic domain. Although the SP algorithm scored significantly worse than the unweighted mean, the SP algorithm in fact forecasted more events correctly in terms of mean percentage accuracy. Both the unweighted mean and confidence-weighted algorithm were less accurate than the SP algorithm by 1.82% (95% CI: [-0.25, 3.78]). This indicates that the weights used by the algorithm were in fact highly effective at identifying experts in the crowd, however, forecasters’ votes, which were being weighted by the SP algorithm, were miscalibrated and not suited for use in the probabilistic domain. In the following section, we propose a version of the SP algorithm adapted to probabilistic forecasting.

3.3 The Meta-Vote Weighting Algorithm

In the current section, we use same theoretical framework used by Wilkening et al. (2020) (see section 6.1) to develop and validate a novel extension of the SP algorithm that allows for the generation of probabilistic forecasts. Our proposed algorithm uses the same weighting function as the SP algorithm, which we identified in Chapter 2. While both the original SP algorithm and our reformulation of the SP algorithm in Chapter 2 have been developed to generate only categorical predictions, our theoretical framework allows us to directly extend the SP algorithm to generate probability forecasts that account for the uncertainty in forecasters' predictions.

In Wilkening et al. (2020), we develop a theoretical framework where forecasters receive signals varying in informativeness. Expert and novice forecasters are distinguished by the potential sets of signals that they receive, and experts receive more-informative signals than novices. We show that under reasonable assumptions, an expert's *contribution* to the aggregate forecast (i.e., the normalised absolute difference between an expert's vote and their meta-prediction) will, on average, always exceed a novice's contribution. Crucially, while our theoretical model shows that experts' and novices' contributions are ordered, such that experts' contributions on average exceed that of novices, it does not prescribe how experts' and novices' predictions should be aggregated.

In chapter 2, we showed that the SP algorithm can be reformulated as a weighted combination of forecasters' votes, weighted by the absolute difference between their votes and meta-predictions:

$$T_{SP}(X_k) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N_k} \frac{|V_{i,k} - M_{i,k}^V| V_{i,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Here, we propose a Meta-Vote Weighting algorithm that uses the absolute difference between forecasters' votes and forecasters' meta-predictions about the proportion of others forecasters voting "true" in the same way to weight forecasters' probability forecasts (i.e., rather than to weight forecasters' votes):

$$T_{MVW}(X_k) = \sum_{i=1}^{N_k} \frac{|V_{i,k} - M_{i,k}^V| P_{i,k}}{\sum_{j=1}^{N_k} |V_{j,k} - M_{j,k}^V|} \quad (3.2)$$

As the Meta-Vote Weighting algorithm proposed here uses the same weights as our reformulation of the SP algorithm, it has the same theoretical properties in terms of being able to identify and leverage experts’ forecasts. By applying the weight to forecasters’ probability forecasts – rather than their votes – we expect this algorithm to be able to generate more accurate aggregated probability forecasts than the original SP algorithm as well as the unweighted mean. In the section below, we test this empirically.

3.4 Experiment 2: Validating the Meta-Vote Weighting

Algorithm

In this section, we validate the Meta-Vote Weighting algorithm against the original SP algorithm and other probabilistic algorithms using the same set of questions from Experiment 1. In the Appendix (Section 6.1), we showed that the weights used by the SP algorithm (and thus the Meta-Vote Weighting algorithm) can be used to identify and leverage expertise in the crowd when theoretical assumptions about the information available to forecasters, and the difficulty of the forecasting problem are met (see Wilkening et al., 2020). A question remains as to whether the Meta-Vote Weighting algorithm will perform well in practice.

The type of forecasting problem at hand is likely to determine whether the assumptions of the model will be met. In order for the Meta-Vote Weighting algorithm to be useful on novel problems where the properties of the forecasting problem is unknown, it needs to exhibit a sufficient degree of robustness when its assumptions are violated. In the previous chapter, we examined forecasting problems specifically where we would expect these assumptions to be violated – for example, on questions that were extremely easy or extremely hard where crowd expertise is likely to be homogeneous. Our results showed that the SP algorithm was fairly robust in most of these cases, such that it did not perform significantly worse even when these assumptions were likely to have been violated. Since the original SP algorithm and the Meta-Vote Weighting algorithm both operate under the same mechanism (i.e., by weighting forecasters by the absolute difference

between their votes and meta-predictions), we would expect the Meta-Vote Weighting algorithm to exhibit a similar degree of robustness against violations of these assumptions. When these assumptions are met, as with datasets where the crowd contains both experts and novices, we would expect the Meta-Vote Weighting algorithm to assign greater weights to high-performing forecasters than low-performing forecasters, and therefore outperform the unweighted mean.

While the Meta-Vote Weighting algorithm is generally expected to outperform the unweighted mean for the reasons discussed above, it is less clear how well the algorithm would perform relative to other probabilistic algorithms. The Meta-Vote Weighting algorithm works by weighting forecasters according to their expertise, as captured by the absolute difference between forecasters' votes and meta-predictions. Due to the fact that it is a linear combination of probability forecasts, the Meta-Vote Weighting algorithm is still likely to be miscalibrated (Ranjan & Gneiting, 2010). In order to correct for miscalibration, we will apply the recalibration function used in Experiment 1 (see Baron et al., 2014; Turner et al., 2014) to the Meta-Vote Weighting algorithm. We predict that this recalibration function would improve the performance of the Meta-Vote Weighting algorithm. In particular, it should outperform the other probabilistic algorithms tested in Experiment 1 – particularly the Minimal Pivoting model, which was the best-performing algorithm overall (refer again to Figure 3.1). To provide an equitable comparison with the other probabilistic algorithms, we also allow for the fact that the Minimal Pivoting model and the P''_{cs} aggregator may also be miscalibrated, and apply the same recalibration function to the forecasts of these algorithms. While we would not necessarily expect these other algorithms to benefit from recalibration as much as the Meta-Vote Weighting algorithm, this approach allows us to identify whether the improvement in score offered by the Recalibrated Meta-Vote Weighting algorithm is due solely to recalibration or more likely due to an interaction between the effects of recalibration and the Meta-Vote Weighting approach. This will allow us to separate the effects of aggregation and recalibration and indicate the extent to which recalibration can be combined effectively with these aggregation approaches.

3.4.1 Methods

Our analyses were conducted on the same dataset as our previous analyses. We compared the forecasting algorithms on the 1196 events, collected across 12 datasets – five of which were provided by Prelec et al. (2017), one large dataset by Lee et al. (2018), and six we collected ourselves. Table 3.1 provides details on the relevant datasets.

We compared the overall performance of the original SP algorithm, four probabilistic algorithms: the Meta-Vote Weighting algorithm, the unweighted mean, the P''_{cs} aggregator, the Minimal Pivoting model, and the recalibrated versions of each of these four probabilistic algorithms. Tables 3.2 and 3.3 provide details on the relevant algorithms. We computed the 95% CIs for the bootstrap mean difference in Brier scores between the Meta-Vote Weighting algorithm and each other algorithm, as well as the mean difference in score between the Recalibrated Meta-Vote Weighting algorithm and each other algorithm. We also generated calibration curves for the top-performing algorithms as we did for Experiment 1, in order to identify the extent of miscalibration for each algorithm.

Additionally, we examined the performance of each algorithm at the dataset level, in order to identify environments best suited to each algorithm.

3.4.2 Results

Figure 3.3 shows the mean transformed Brier score of the Surprisingly popular (SP), unweighted mean (UM), Recalibrated unweighted mean (UM-R), P''_{cs} , Recalibrated P''_{cs} aggregator ($P''_{cs} - R$), Minimal Pivoting (MP), recalibrated Minimal Pivoting (MP-R), Meta-Vote Weighting (MVW), and Recalibrated Meta-Vote Weighting (MVW-R) algorithms across all 1196 events. Below, we report the mean difference in score and the bootstrap 95% CI for mean paired difference between (1) the Meta-Vote Weighting algorithm and each other algorithm, and (2) between the Recalibrated Meta-Vote Weighting algorithm and each other algorithm.

Performance of the Meta-Vote Weighting algorithm

As predicted, the Meta-Vote Weighting algorithm (rightmost dark grey bar in Figure 3.3) provided significantly better forecasts than the original SP algorithm. However, the Meta-Vote Weighting

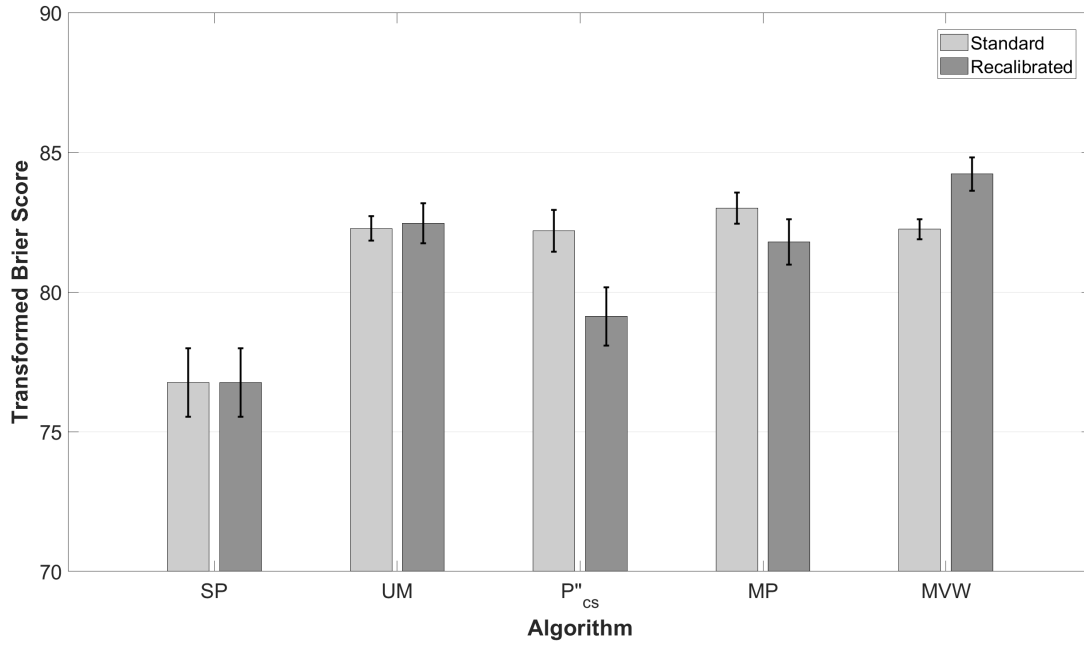


Figure 3.3: The mean transformed Brier score for the Surprisingly Popular (SP) unweighted mean (UM), P''_{cs} , Minimal Pivoting (MP), and Meta-Vote Weighting (MVW) algorithms over 1196 events across 12 datasets (light grey), along with the recalibrated version of each algorithm (dark grey). Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms, recalibrated or otherwise.

algorithm offered little to no advantage over the other standard probabilistic algorithms. It scored significantly better than the original SP algorithm ³ by 5.49 points (95% CI: [3.58, 7.48]) and the Recalibrated P''_{cs} aggregator by 3.12 points (95% CI: [1.58, 4.81]); significantly worse than the Minimal Pivoting model by -0.75 points (95% CI: [-1.24, -0.23]) and the Recalibrated Meta-Vote Weighting algorithm by -1.97 points (95% CI: [-2.49, -1.40]); and neither significantly better nor worse than the unweighted mean (95% CI: [-0.40, 0.35]), the Recalibrated unweighted mean (95% CI: [-1.08, 0.71]), the P''_{cs} aggregator (95% CI: [-0.93, 1.15]), or the Recalibrated Minimal Pivoting model (95% CI: [-0.57, 1.55]).

Performance of the Recalibrated Meta-Vote Weighting algorithm

As Figure 3.3 suggests, the Recalibrated Meta-Vote Weighting algorithm (the rightmost light-blue bar) significantly outperformed all other algorithms, including the Minimal Pivoting model and the Meta-Vote Weighting algorithm. It outperformed the original SP algorithm by 7.46 points (95% CI: [5.90, 9.12]), the unweighted mean by 1.95 points (95% CI: [1.34, 2.54]), the Recalibrated unweighted mean by 1.76 points (95% CI: [1.01, 2.44]), the P''_{cs} aggregator by 2.03 points (95% CI: [1.20, 3.08]), the Recalibrated P''_{cs} aggregator by 5.09 points (95% CI: [3.69, 6.37]), the Minimal Pivoting model by 1.22 points (95% CI: [0.80, 1.60]), the Recalibrated Minimal Pivoting model by 2.43 points (95% CI: [1.80, 3.13]), and the Meta-Vote Weighting algorithm by 1.97 points (95% CI: [2.49, 1.40]).

Calibration curves

To obtain better insight into why the Meta-Vote Weighting algorithm performed much worse than we expected and why extremisation was effective for the Meta-Vote Weighting algorithm but not the other algorithms, we generated calibration curves for the six top-performing algorithms: the unweighted mean, the P''_{cs} aggregator, the Minimal Pivoting model, the Meta-Vote Weighting algorithm, and the Recalibrated Meta-Vote Weighting algorithm.

Figure 3.4 shows the calibration curves for the six top-performing algorithms. The Meta-Vote Weighting algorithm (the light-blue line) appears to be the worst-calibrated algorithm in

³Note that the recalibrated SP algorithm makes identical forecasts to the original SP algorithm, and thus both algorithm have the same level of performance all forecasts.

almost every bin. The miscalibration appears to be systematic in that the Meta-Vote Weighting algorithm consistently under-estimates the probability of events that are “true” by assigning lower probability to those events than it should. For example, it is most miscalibrated at the 60–70 percentile bin, where it forecasts approximately 93% events correctly but only assigns probabilities between .6 to .7 to those events. A similar but weaker effect can be observed for the surrounding bins, and to a lesser extent, the inverse effect for events where the outcome is “false”. This pattern of miscalibration illustrates why the extremisation approach is particularly effective for the Meta-Vote Weighting algorithm, but less so for the other algorithms. Extremisation shifts the responses in the percentiles close to .5, where there is high uncertainty, closer to the ends of the scale. As the Meta-Vote Weighting algorithm is more miscalibrated than other algorithms in these bins where extremisation would be most effective, extremisation should thus offer a greater increase in score for the Meta-Vote Weighting algorithm than for any of the other algorithms. Indeed, we can see from these calibration curves that the Recalibrated Meta-Vote Weighting algorithm (dark blue) is extremely well-calibrated – it is almost perfectly calibrated in the 20–30, 30–40, 50–60, 60–70, and 90–100 percentile bins, and is only slightly miscalibrated in the remaining bins.

It is important to note that while the improvement in score offered by the Recalibrated Meta-Vote Weighting algorithm over other algorithms seems small, it is in fact quite substantial. If we examine the increase in score over the unweighted mean offered by the Minimal Pivoting model and the Recalibrated unweighted mean, these differences were 0.27 points and 0.18 points respectively. In contrast, the improvement offered by the Recalibrated Meta-Vote Weighting algorithm over the unweighted mean was 1.95 points. The Recalibrated Meta-Vote Weighting algorithm therefore produces an increase in score over 7 times more than the next-best algorithm (the Minimal Pivoting model). Similarly, the Recalibrated Meta-Vote Weighting algorithm outperforms the Minimal Pivoting model by 1.22 points, about 4.5 times the improvement offered by the Minimal Pivoting model over the unweighted mean. The Recalibrated Meta-Vote Weighting algorithm thus provides an improvement in forecasting performance superior to existing probabilistic forecasting algorithms in the single-question domain.

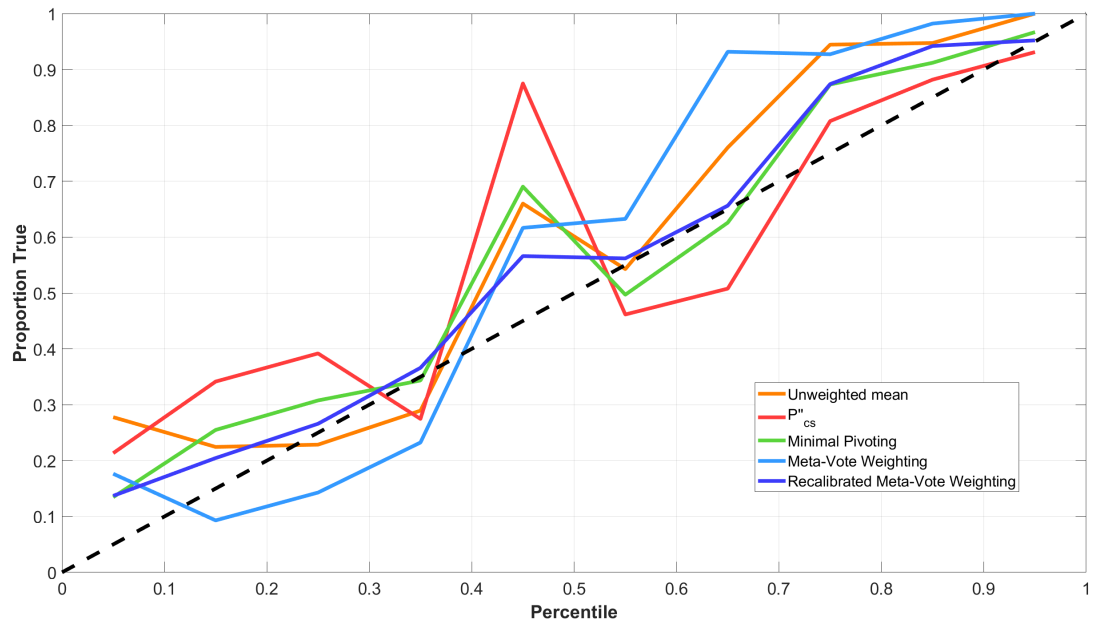


Figure 3.4: The calibration curve for the unweighted mean (orange), the P''_{cs} aggregator (red), the Minimal Pivoting model (light green), the Meta-Vote Weighting algorithm (light blue), and the Recalibrated Meta-Vote Weighting algorithm (dark blue). Each calibration curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the calibration of a perfectly-calibrated forecaster (or algorithm), whose forecasted probabilities match exactly the proportion of “true” events observed.

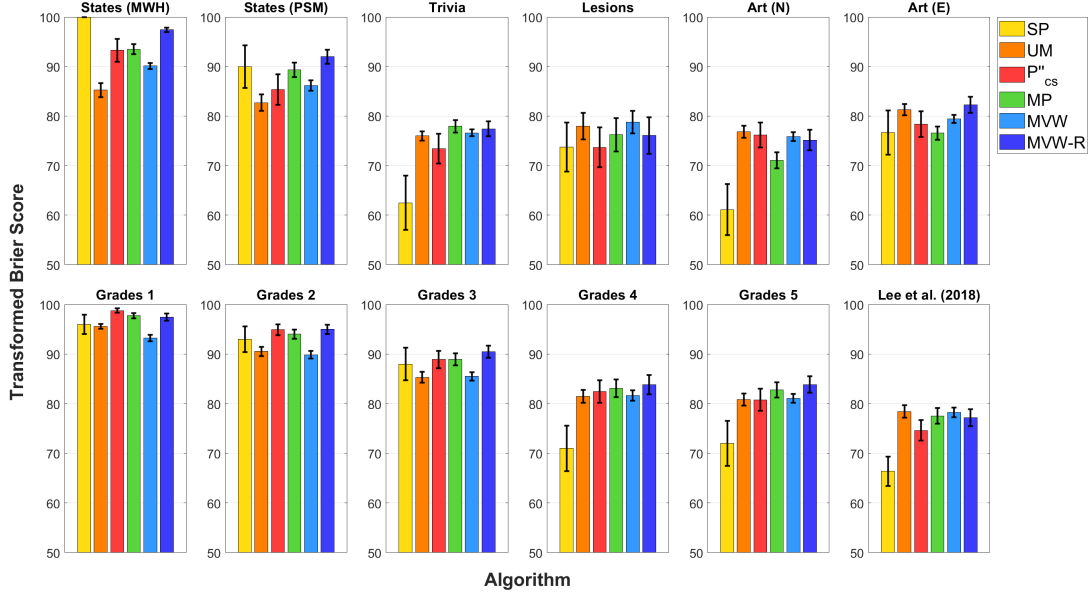


Figure 3.5: The mean transformed Brier score for the Surprisingly Popular algorithm (yellow), the unweighted mean (orange), the P''_{cs} aggregator (red), the Minimal Pivoting model (light green), the Meta-Vote Weighting algorithm (light blue), and the Recalibrated Meta-Vote Weighting algorithm (dark blue) on each individual dataset. Error bars show the standard error.

Performance at the dataset level

To obtain deeper insight into the datasets where these algorithms offered the greatest improvement, we compared the performance of the Recalibrated Meta-Vote Weighting algorithm and the other algorithms at the dataset level. Figure 3.5 shows the mean transformed Brier score of each algorithm separately for each dataset.

In Figure 3.5, we see that the Meta-Vote Weighting algorithm (light blue bar) offers an advantage over the original SP algorithm (yellow bar) for the majority of datasets, in particular, for datasets with high difficulty where most algorithms tend to perform poorly. This is not surprising, given that forecasters' probability forecasts contain important information about the uncertainty in the outcome, and the Meta-Vote Weighting algorithm generates predictions using a weighted average of these forecasts. In contrast, the original SP algorithm, which only uses forecasters' votes, does not make use of the information contained in forecasters' probability forecasts at all, and therefore performed poorly on datasets where there were many events with

uncertain outcomes (e.g., the highest difficulty Grades dataset).

Contrary to our theoretical predictions, but consistent with the results seen in the other figures, the Meta-Vote Weighting algorithm (light blue bar in Figure 3.5) offered little to no advantage over the unweighted mean (orange bar) on almost every dataset. The Meta-Vote Weighting algorithm did not consistently outperform either the P''_{cs} aggregator (red bar) or Minimal Pivoting model (light green bar) either. Referring back to the calibration plot for each algorithm (Figure 3.4), we can see that this is likely due to the extent to which the Meta-Vote Weighting algorithm is much more miscalibrated than the other algorithms. The recalibration function, when applied to the Meta-Vote Weighting algorithm, was able to correct for this miscalibration and improve predictions for nine out of twelve datasets, often by a substantial amount. Although the Recalibrated Meta-Vote Weighting algorithm did not offer the most accurate set of predictions on every single dataset, it performed consistently well for a majority of these datasets. The Recalibrated Meta-Vote Weighting algorithm was therefore robust in that it performed well even when the assumptions about heterogeneity in forecaster expertise and task difficulty were likely to have been violated, such as in the Grades 4 and 5 datasets.

More importantly, our results show that neither the Meta-Vote Weighting algorithm nor recalibration alone (e.g., when applied to the unweighted mean) was sufficient to significantly outperform other probabilistic forecasting algorithms. However, when these approaches were combined, the resulting algorithm was able to produce significantly more accurate predictions than any other probabilistic algorithm, including algorithms that have also been recalibrated. This approach of combining recalibration with the Meta-Vote Weighting algorithm is therefore an effective way of generating accurate probabilistic forecasts in the single-question domain and, importantly, seems to outperform other existing aggregation approaches.

Before we discuss the implications of these findings, we first report an additional, post-hoc analysis to address an alternative explanation for these results.

3.4.3 Optimally recalibrating model predictions

One possible explanation for why the Recalibrated Meta-Vote Weighting algorithm outperformed the recalibrated versions of the unweighted mean, the P''_{cs} aggregator, and Minimal Pivoting

model is that the parameters of the recalibration function we applied (see Baron et al., 2014) is better-suited for the Meta-Vote Weighting algorithm than for these other algorithms. In the current section we examine whether these other algorithms could outperform the Recalibrated Meta-Vote Weighting algorithm if we apply the recalibration function with optimally-estimated parameter values.

To provide an advantageous comparison for these other algorithms, we use leave-one-out cross-validation (i.e., a jack-knife)⁴ to estimate the optimal parameters in the recalibration function that we are applying to each algorithm, excluding the Meta-Vote Weighting algorithm. We note that, although cross-validation is the most common approach in the literature for applying these recalibration functions, using cross-validation to estimate these parameters is not feasible in the single-question domain because it requires records of forecasters’ past performance on questions with known outcomes. Nonetheless, we adopt this approach here since it allows us to identify the upper-limit on improvement from the recalibration function when it is applied to these algorithms. Allowing the parameter values in the recalibration function to vary also captures if these algorithms need to be anti-extremised instead, rather than extremised. If the difference in score between these algorithms could be accounted for predominantly by the parameters of the recalibration function, then we should expect to see a substantial increase in score for the recalibrated versions of these algorithms, such that the Recalibrated Meta-Vote Weighting algorithm would no longer significantly outperform these other algorithms. On the other hand, if the increase in score is small, such that the Recalibrated Meta-Vote Weighting algorithm still outperforms the other optimally recalibrated approaches, then this would indicate that our model provides a unique improvement in forecasting performance above and beyond that of recalibration combined with any of the other algorithms.

For each of the 1196 events in the combined dataset, we used a leave-one-out cross-validation approach to estimate the optimal recalibration parameter (See Equation 3.1) for the aggregated forecast from each of the unweighted mean, the P''_{cs} aggregator, and the Minimal Pivoting model. For each algorithm, we tested values for the a parameter ranging from 0 to 10 in increments of

⁴We used leave-one-out cross-validation since it appears to be the most common cross-validation approach in the literature. Nonetheless, we note that the use of leave-one-out cross-validation over alternative cross-validation procedures remains a hotly discussed topic in both the machine learning and cognitive modelling literature (e.g., Fushiki, 2011; Gronau & Wagenmakers, 2019).

0.1 on each training set, comprised of $K - 1$ events, where K is total number of events. The test question on any given iteration was never included in the training set. We then selected the a value that led to the highest average mean score across all training events and applied the recalibration function with that a parameter value to the aggregated forecast for that test event. We repeated this for each event and for each algorithm, and then compared the average mean transformed Brier score of each of these recalibrated approaches against the score of the Recalibrated Meta-Vote Weighting algorithm, where the recalibration parameter was not optimised, but fixed at 2.5. We then computed bootstrap 95% confidence intervals for the mean paired difference in score between the Recalibrated Meta-Vote Weighting algorithm and the three other recalibrated approaches, and we examined whether any of these differences were statistically significant.

Figure 3.6 shows the mean score for the Recalibrated unweighted mean, Recalibrated P''_{cs} aggregator, Recalibrated Minimal Pivoting model, and Recalibrated Meta-Vote Weighting algorithm. The Recalibrated Meta-Vote Weighting algorithm significantly outperformed the optimally recalibrated unweighted mean by 1.36 points (95% CI: [0.82, 1.90]), the optimally recalibrated P''_{cs} aggregator by 1.81 points (95% CI: [1.01, 2.69]), and the optimally recalibrated Minimal Pivoting model by 1.19 points (95% CI: [0.79, 1.58]). Surprisingly, the Recalibrated Meta-Vote Weighting algorithm still significantly outperformed each other algorithm, even when those other algorithms' forecasts were optimally recalibrated. These results show that even if the answers to a large set of training questions were hypothetically available and each algorithm could be optimally recalibrated using this information, there would be recalibration parameter that would allow these other aggregation approaches to outperform the Recalibrated Meta-Vote Weighting algorithm (for which the parameter in the recalibration function was fixed at 2.5, and thus not estimated from the data). The Recalibrated Meta-Vote Weighting algorithm therefore appears to generate uniquely superior predictions to the other existing single-question probability forecasting algorithms in the literature.

3.4.4 Discussion

In the results thus far, we have developed and validated a novel algorithm for aggregating probability forecasts on single-question forecasting problems – forecasting problems where forecasters'

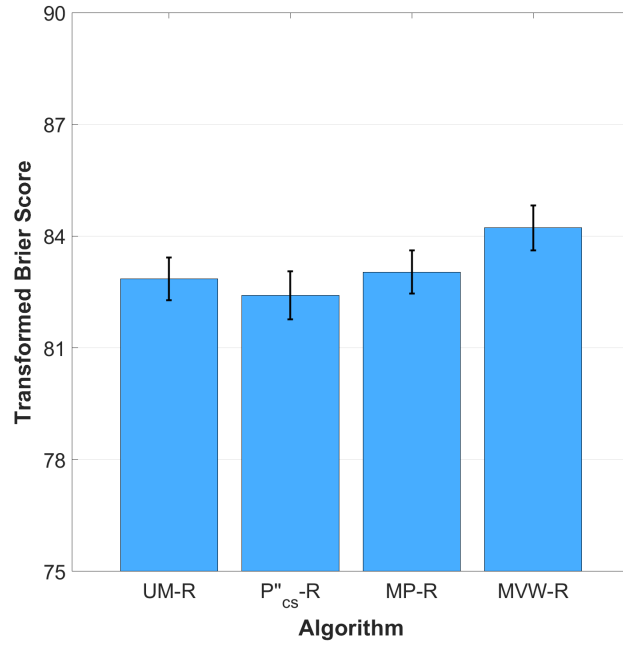


Figure 3.6: The mean transformed Brier score for the optimally recalibrated versions of the unweighted mean (UM), the P''_{cs} aggregator, and the Minimal Pivoting (MP) algorithm, compared to the Recalibrated Meta-Vote Weighting (MVW) algorithm, which has not been optimally recalibrated. Error bars show the standard error. The Recalibrated Meta-Vote Weighting algorithm (far right) still significantly outperforms all other algorithms.

responses to only one question of interest are available. Our proposed algorithm, the Recalibrated Meta-Vote Weighting algorithm, combines the forecast-recalibration approach (e.g., Baron et al., 2014) with a novel extension of the Surprisingly Popular algorithm (Prelec et al., 2017) that generates better predictions in the probabilistic domain. We sought to find the best-performing aggregation algorithm overall across a wide range of single-question forecasting problems. To achieve this, we tested and compared the performance of the Recalibrated Meta-Vote Weighting algorithm against the performance of standard probabilistic-forecasting algorithms: the unweighted mean, the Recalibrated unweighted mean (e.g., Baron et al., 2014; Turner et al., 2014), the P''_{cs} aggregator (Satopää et al., 2016), and the Minimal Pivoting model (Palley & Soll, 2019). We assessed the performance of these algorithms using a transformed Brier score (e.g., Budescu & Chen, 2015) on a large dataset of 1196 forecasting problems combined from 12 smaller datasets: five datasets provided by Prelec et al. (2017), six datasets that we collected ourselves and had used in Chapter 2, and one dataset provided by Lee et al. (2018).

Our results showed that the Recalibrated Meta-Vote Weighting algorithm outperformed all other algorithms by a substantial amount. The improvement in Brier score offered by the Recalibrated Meta-Vote Weighting algorithm over both the unweighted mean and other probabilistic algorithms was more than four times greater than the improvement offered by any other algorithm over the unweighted mean. Even when we allowed for other algorithms to be optimally recalibrated using training data, the Recalibrated Meta-Vote Weighting algorithm still significantly outperformed all other algorithms, recalibrated or otherwise. Although the Recalibrated Meta-Vote Weighting algorithm did not outperform every other algorithm on each individual dataset, it was among the top-performing algorithms for almost every dataset, making it extremely effective and also highly robust, performing well across a wide and diverse range of forecasting environments.

Contributions of the present work

Our results have important theoretical and practical implications for how decision analysts should aggregate probability forecasts in the single-question domain. The impressive performance of the Recalibrated Meta-Vote Weighting algorithm provides a novel demonstration of how expertise can be identified and leveraged using the normalised difference between forecasters' votes and

meta-predictions in the probabilistic domain. While the efficacy of this weighting measure at identifying expertise was demonstrated comprehensively in the previous chapter, the current chapter provides compelling evidence that this measure of expertise can also be used to leverage expertise effectively in the probabilistic domain. Although the Meta-Vote Weighting algorithm could identify and weight experts appropriately, experts' forecasts appeared to be consistently under-confident, leading to miscalibration in the aggregated forecast provided by the Meta-Vote Weighting algorithm. Once we corrected for this bias using a simple recalibration function, the Meta-Vote Weighting algorithm was able to successfully leverage the forecasts of experts in the crowd in order to produce more accurate probability forecasts.

Empirically, our results suggest that the optimal aggregation approach for novel single-question forecasting problems would be to apply the Recalibrated Meta-Vote Weighting algorithm. The robust performance of this algorithm across a wide and diverse range of environments (e.g., Figure 3.5) suggests that it would likely outperform other probabilistic-forecasting algorithms on novel forecasting problems in the single-question domain. In particular, our results indicate that the Recalibrated Meta-Vote Weighting algorithm is likely to perform well relative to other aggregation approaches, even when there is no history of forecasters' past performance, no other external information about the expertise of forecasters, nor any information about the properties of the forecasting problem (e.g., whether there is substantial uncertainty in the outcome). This algorithm thus provides a useful and unique forecasting solution in the single-question probabilistic-forecasting domain, where few effective aggregation approaches have been developed to date.

Relationship to previous work

The Recalibrated Meta-Vote Weighting algorithm was synthesised from a combination of ideas in the forecasting literature. The Meta-Vote Weighting algorithm shares the same mechanism with the SP algorithm in being able to identify and leverage experts' forecasts over novices' because it was developed using the same theoretical framework from Chapter 2 (see section 6.1 for details). The original SP algorithm generated poor probabilistic predictions for many of these datasets (such as on NFL predictions; Lee et al., 2018) where outcomes had high uncertainty, partly because the original SP algorithm could only produce binary forecasts, and partly because it did not aggregate the uncertainty (or confidence) from forecasters' votes. In contrast, the Meta-Vote

Weighting algorithm is able to generate more-accurate predictions by using forecasters’ probability forecasts, which contain information about their perceived uncertainty in each outcome. Although the unweighted mean, P''_{cs} aggregator and the Minimal Pivoting model also aggregate forecasters’ probability forecasts, they operate under different mechanisms that do not seek to leverage expertise in the same way as the Meta-Vote Weighting algorithm. Our results have provided the first empirical demonstration in the probabilistic forecasting domain that forecasters’ expertise can be identified using the normalised difference between their votes and meta-predictions.

The finding that the Recalibrated Meta-Vote Weighting algorithm significantly outperformed all other algorithms, even when those algorithms were recalibrated using training data, is largely consistent with the existing findings from the literature. Ranjan and Gneiting (2010) showed that algorithms that linearly combine forecasts are necessarily miscalibrated and need extremisation. The Meta-Vote Weighting algorithm would be expected to be more miscalibrated than the P''_{cs} aggregator and the Minimal Pivoting model, since the Meta-Vote Weighting algorithm weights forecasters’ probabilistic predictions by their normalised contributions, whereas these other algorithms do not weight forecasters’ predictions at all. The Meta-Vote Weighting algorithm was therefore the algorithm that was likely to benefit the most from recalibration, as we observed in our results. Indeed, we could also see from the calibration curves in Figure 3.4 that the Meta-Vote Weighting algorithm and the unweighted mean (both algorithms that linearly weight forecasters) were more miscalibrated than the Minimal Pivoting model, and to a lesser extent, the P''_{cs} aggregator. Our results therefore appear to be consistent in this regard with the theoretical predictions from previous work.

Extensions to the current work

In the sections above, we have developed a natural extension of the SP algorithm to probabilistic forecasting using the theoretical framework developed in Wilkenning et al. (2020). Our theoretical framework showed that under reasonable assumptions, the normalised absolute difference between forecasters’ votes and their meta-predictions about the proportion of others voting “true” is strictly larger for experts than novices, and therefore can be used to identify and assign experts’ predictions greater weight compared to novices’ predictions. Although this measure was effective for identifying expertise in both the categorical and probabilistic forecasting domains, it requires

relatively strong theoretical assumptions about the heterogeneity of forecasters' expertise, the information available to forecasters, and the difficulty of the forecasting problem. Theoretically, forecasters' expertise can be identified much more effectively, without requiring such strong assumptions, by using a probabilistic version of the normalised difference instead. The following experiment looks to develop and test the efficacy of this proposed novel weighting approach.

3.5 Experiment 3: A Refined Measure of Crowd Expertise

In this section, we propose a novel weighting algorithm, the Meta-Probability Weighting algorithm, that weights forecasters' probability forecasts by the normalised absolute difference between their probability forecasts and their meta-prediction about the average probability forecast by others, formally:

$$T_{MPW}(X_k) = \sum_{i=1}^{N_k} \frac{|P_{i,k} - M_{i,k}^P| P_{i,k}}{\sum_{j=1}^{N_k} |P_{j,k} - M_{j,k}^P|}$$

where $P_{i,k}$ is the probability forecast of the i th forecaster for the k th event, and $M_{i,k}^P$ is their meta-prediction about the average probability forecast of others.

The Meta-Probability Weighting algorithm operates under the same theoretical principles as the SP algorithm discussed in Chapter 2. Previous findings from the social psychology literature have shown that experts, compared with non-experts, tend to have better knowledge about the true probability of outcomes, better meta-knowledge about the expertise of others, and better meta-knowledge about their own expertise relative to others' expertise (e.g., see Dunning et al., 2004, 1989; Krueger, 1998; Kruger & Dunning, 1999). In contrast, novices lack both knowledge about the probability of outcomes and meta-knowledge about their expertise relative to the expertise of others. In the forecasting context, experts are therefore able to identify when other forecasters are likely to be novices for a particular question and recognise that the crowd will predict differently from themselves. In contrast, novices will not expect others' forecasts to differ substantially from their own, since they lack meta-cognitive awareness about the expertise of others and therefore what others will predict. Thus, regardless of whether predictions are measured in the form of votes or probabilities, or whether meta-predictions are measured as the

average vote or probability forecasted by others, the absolute difference between a forecasters’ predictions and meta-predictions about the forecasts of others should, in theory, correspond to differences in forecasters’ expertise.

In the previous chapter and the first two experiments in this chapter, we have provided ample evidence that the first metric – the absolute difference between forecasters’ votes and meta-predictions about the proportion of others voting “true” – is predictive of expertise and therefore can be leveraged to generate accurate predictions in the single-question domain. In the remainder of this chapter, we demonstrate that the second metric – the absolute difference between forecasters’ probability forecasts and meta-predictions about the average probability forecasted by others – provides an even better measure of forecaster expertise than the first metric. In Wilkening et al. (2020), we extend our theoretical framework to show the Meta-Probability Weighting metric is not only an effective measure of expertise, but also distinguishes between experts and novices under more general theoretical conditions. Furthermore, we show that the weights used by the SP algorithm and the Meta-Vote Weighting algorithm require additional, stronger assumptions in order to distinguish between experts and novices (Wilkening et al., 2020).

In theory, the Meta-Probability Weighting metric provides a better measure of expertise than the Meta-Vote Weighting metric used by the SP algorithm due to the responses used for calculating forecasters’ weights (i.e., forecasters’ votes vs. forecasters’ probabilistic forecasts). Forecasters’ votes are categorical and can only take one of two possible values – as such, novices can only be distinguished from experts when each group votes for a different outcome. In contrast, probability forecasts, which are continuous variable, are able to capture small differences in expertise (and thus probability forecasts) between forecasters. For example, forecasters who predict the correct outcome with certainty can be distinguished from forecasters who may have guessed the correct answer by chance, whereas it is not possible to distinguish between these two forecasters by their votes. For this reason, the weights of the Meta-Probability Weighting algorithm should provide a better measure of expertise than the weights used by the Meta-Vote Weighting algorithm.

An empirical question remains as to whether the Meta-Probability Weighting algorithm will perform well in practice. One challenge in testing these aggregation approaches empirically is that the current datasets available in the literature do not allow for an equitable comparison

between these algorithms. There are no studies to date that have collected both forecasters’ meta-predictions about the proportion of others voting “true” and forecasters’ meta-predictions about the average probability forecasted by others. As a preliminary test of these algorithms, we compared the performance of each algorithm using the datasets we collected thus far, prior to collecting a new set of responses to both meta-prediction questions. As these datasets only contained forecasters’ meta-predictions about the proportion of others voting “true”, we made the assumption that forecasters’ meta-predictions about the average probability forecasted by others would be identical to those responses. While this would have most likely disadvantaged the Meta-Probability Weighting algorithm against the Meta-Vote Weighting algorithm, these results would at least provide some indication of how well the Meta-Probability Weighting algorithm would perform with the appropriate meta-predictions from forecasters. For the same reasons that we recalibrated the Meta-Vote Weighting algorithm (Ranjan & Gneiting, 2010), we applied the same recalibration function and fixed parameter from Baron et al. (2014) to the Meta-Probability Weighting algorithm to test whether synthesising these approaches would result in better forecasts.

3.5.1 Methods

We conducted the same analysis as above in order to examine the performance of the Meta-Probability Weighting algorithm and the Recalibrated Meta-Probability Weighting algorithm. Specifically, we compared the performance of the Meta-Probability Weighting algorithm and the Recalibrated Meta-Probability Weighting algorithm to that of the unweighted mean, the P''_{cs} aggregator, the Minimal Pivoting model, the Meta-Vote Weighting algorithm, and the Recalibrated Meta-Vote Weighting algorithm on the same set of questions as before. We computed the 95% CIs for the bootstrap mean difference in Brier score between (1) the Meta-Probability Weighting algorithm and each other algorithm, and (2) the Recalibrated Meta-Probability Weighting algorithm and each other algorithm. We also generated calibration curves for the Meta-Probability Weighting algorithm and the Recalibrated Meta-Probability Weighting algorithm and compared them to the calibration curves for the Meta-Vote Weighting algorithm and Recalibrated Meta-Vote Weighting algorithm. As before, we also compared the performance of algorithms at the dataset level in order to identify whether the differences in performance were generalised or specific to

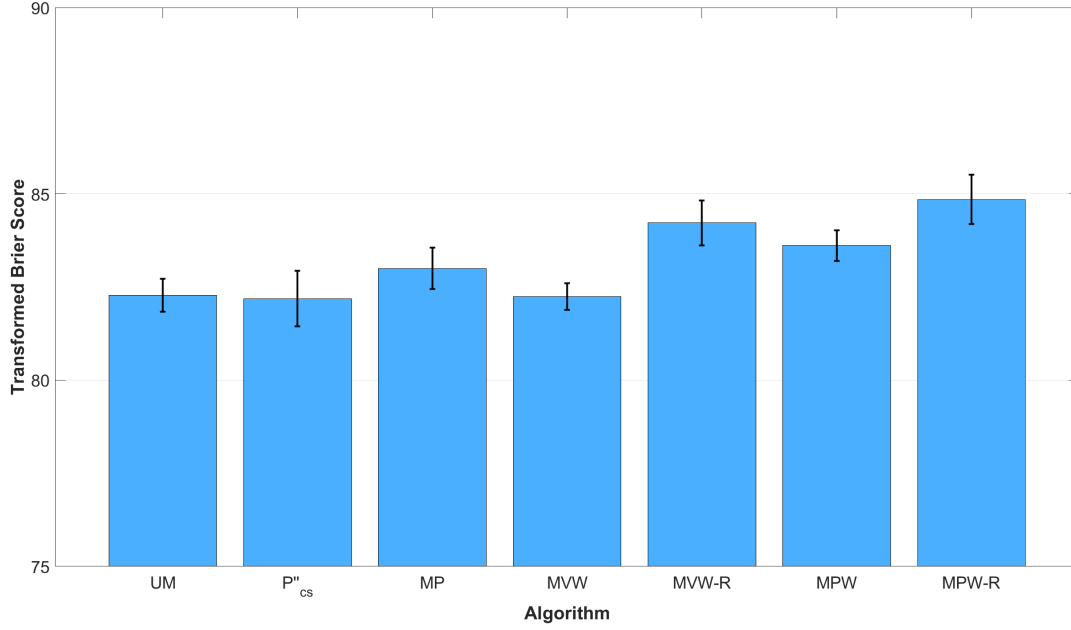


Figure 3.7: The mean transformed Brier score for the unweighted mean (UM), the P''_{cs} aggregator, the Minimal Pivoting (MP) algorithm, the Meta-Vote Weighting (MVW) algorithm, the Recalibrated Meta-Vote Weighting (MVW-R) algorithm, the Meta-Probability Weighting algorithm (MPW), and the Recalibrated Meta-Probability Weighting (MPW-R) algorithm over 1196 events across 12 datasets. Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms.

certain datasets, and therefore particular forecasting environments.

3.5.2 Results

Figure 3.7 shows the mean transformed Brier score for each algorithm across the 1196 events across the 12 datasets.

Performance of the Meta-Probability Weighting algorithm

The basic Meta-Probability Weighting algorithm provided significantly better forecasts than the unweighted mean by 1.34 points (95% CI: [0.94, 1.74]), the P''_{cs} aggregator by 1.42 points (95% CI: [0.51, 2.46]), the Minimal Pivoting model by 0.61 points (95% CI: [0.11, 1.13]), and the Meta-Vote Weighting algorithm by 1.36 points (95% CI: [1.14, 1.58]), but provided significantly

worse forecasts than the Recalibrated Meta-Vote Weighting algorithm by -0.61 points (95% CI: [-1.07, -0.07]) and the Recalibrated Meta-Probability Weighting algorithm by -1.24 points (95% CI: [-0.64, -1.79]). The Meta-Probability Weighting algorithm therefore significantly outperformed every other algorithm that was not recalibrated, including the Meta-Vote Weighting algorithm proposed earlier in this chapter.

Performance of the Recalibrated Meta-Probability Weighting algorithm

Recalibration provided a large improvement to the score of the Meta-Probability Weighting algorithm, and the Recalibrated Meta-Probability Weighting algorithm significantly outperformed all other aggregation approaches, including the previous best-performing algorithm, the Recalibrated Meta-Vote Weighting algorithm. Specifically, the Recalibrated Meta-Probability Weighting algorithm significantly outperformed the unweighted mean by 2.58 points (95% CI: [1.85, 3.27]), the P''_{cs} aggregator by 2.66 points (95% CI: [1.81, 3.62]), the Minimal Pivoting model by 1.85 points (95% CI: [1.18, 2.50]), the Meta-Vote Weighting algorithm by 2.60 points (95% CI: [1.82, 3.29]), and the Recalibrated Meta-Vote Weighting algorithm by 0.63 points (95% CI: [0.22, 1.03]).

Calibration curves

Figure 3.8 shows the calibration curves for the Meta-Vote Weighting algorithm, Recalibrated Meta-Vote Weighting algorithm, Meta-Probability Weighting algorithm, and Recalibrated Meta-Probability Weighting algorithm. Comparing the curves for the Meta-Vote Weighting algorithm (orange) and Meta-Probability Weighting algorithm (light green), we can see that the Meta-Probability Weighting algorithm is slightly better calibrated for the 10–20, 20–30, and 40–50 percentile bins. Recalibration appeared to be approximately equally effective for both algorithms. Comparing the curves for the Recalibrated Meta-Vote Weighting algorithm (red) and the Recalibrated Meta-Probability Weighting algorithm (blue), we can see that the Recalibrated Meta-Probability Weighting algorithm is only slightly better-calibrated for the 40–50 and 70–80 percentile bins, but worse for the 20–30 and 30–40 percentile bins. These small improvements in calibration for the Recalibrated Meta-Probability Weighting algorithm was sufficient for it to significantly outperform the Recalibrated Meta-Vote Weighting algorithm by 0.63 points in Brier score.

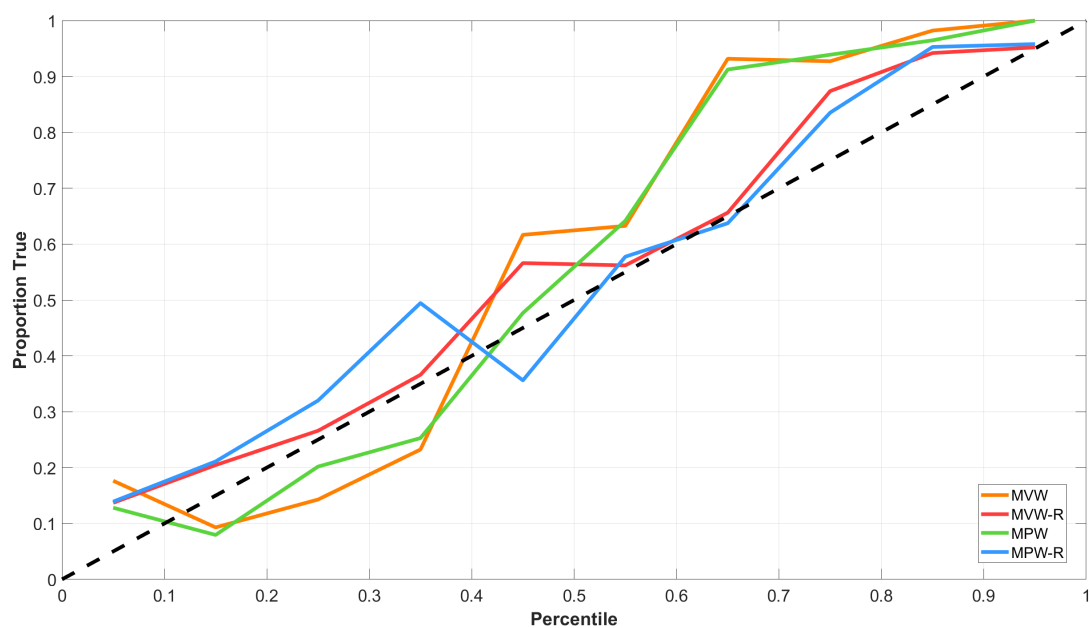


Figure 3.8: The calibration curve for the Meta-Vote Weighting algorithm (orange), the Recalibrated Meta-Vote Weighting algorithm (red), the Meta-Probability Weighting algorithm (light green), and the Recalibrated Meta-Probability Weighting algorithm (blue). Each calibration curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the curve of a perfectly calibrated forecaster (or algorithm), whose forecasted probabilities match exactly the proportion of “true” events observed.

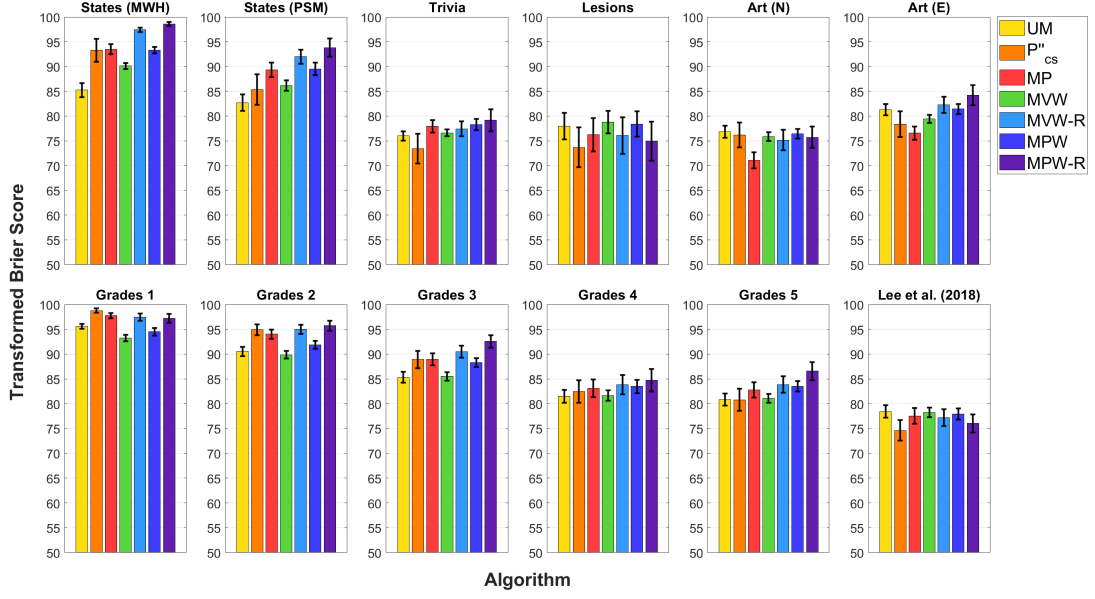


Figure 3.9: The mean transformed Brier score for the unweighted mean (yellow), the P''_{CS} aggregator (orange), the Minimal Pivoting model (red), the Meta-Vote Weighting algorithm (light green), the Recalibrated Meta-Vote Weighting algorithm (light blue), the Meta-Probability Weighting algorithm (dark blue), and the Recalibrated Meta-Probability Weighting algorithm (purple) on each individual dataset. Error bars show the standard error.

Performance at the dataset level

The performance of each algorithm on each individual dataset is shown in Figure 3.9. The Recalibrated Meta-Probability Weighting algorithm (purple bar) appears to perform well in general, providing the best forecast for most of these datasets. It provides the greatest improvement over the unweighted mean (yellow bar) for the two US States datasets and all the Grades datasets except Grades 1 (the easiest difficulty), also outperforming other aggregation approaches, but to a slightly lesser extent. The Recalibrated Meta-Probability Weighting algorithm thus appears to provide the best forecasts out of these standard approaches across a wide variety of forecasting environments. Furthermore, the Recalibrated Meta-Probability Weighting algorithm also appears to be robust in all the datasets, demonstrating no major loss in score compared to the other algorithms for any datasets.

3.5.3 Discussion

The Meta-Probability Weighting algorithm significantly outperformed all other non-recalibrated algorithms, demonstrating that the absolute difference between forecasters' probability forecasts and their meta-predictions about the average probability forecast of others was an effective measure of expertise in the single-question domain. Our results provide the first empirical demonstration in the literature that this measure of expertise can be used to aggregate forecasters' predictions effectively.

Our results also indicate the weights used by the Meta-Probability Weighting algorithm provide an even better measure of expertise than the weights used by the Meta-Vote Weighting algorithm. This is not surprising, since the Meta-Probability Weighting algorithm uses forecasters' probability forecasts, rather than forecasters' votes, to identify expertise. As forecasters' probability forecasts are more effective than forecasters' votes at quantifying expertise, the Meta-Probability Weighting metric therefore allows us to distinguish between expertise in many cases where the Meta-Vote Weighting metric would not. Indeed, our empirical findings are highly consistent with this theoretical advantage of the Meta-Probability Weighting algorithm over the Meta-Vote Weighting algorithm.

Extremisation improved the score of the Meta-Probability Weighting algorithm even further, with the Recalibrated Meta-Probability Weighting algorithm significantly outperforming every other aggregation algorithm. The most accurate probability forecasts could only be produced by combining Meta-Probability Weighting with recalibration.

Our results are limited by the assumption of equivalence between (1) forecasters' meta-predictions about the average probability forecast of others, and (2) forecasters' meta-predictions about the proportion of others voting $\hat{p}_{true}$. The latter was assumed to be an appropriate substitute for the former since only the latter was measured in these datasets. While this means we are likely to be underestimating the performance of the Meta-Probability Weighting algorithm in the same way that we have for the Minimal Pivoting model, the exact performance that these algorithms would have obtained is unknown. In order to correctly gauge the relative efficacy of these algorithms, it is necessary to conduct an experiment whereby forecasters provide responses to both types of meta-prediction questions.

3.6 Experiment 4: Validating the Meta-Probability

Weighting Algorithm

We conducted a replication of Experiment 2 from the previous chapter, which used questions from US grade school that varied systematically in difficulty, using a new sample of forecasters and eliciting both types of meta-prediction questions. By collecting forecasters' responses over a diverse range of questions, we aimed to provide a fairer comparison between the Meta-Probability Weighting algorithm and (1) the Minimal Pivoting model, which requires forecasters' meta-predictions about the average probability forecast of others; (2) the Meta-Vote Weighting algorithm, which require forecasters' meta-predictions about the proportion of others voting "true"; and (3) the unweighted mean and P''_{cs} aggregator, which require only probability forecasts. By using the same experimental design and collecting both types of meta-prediction questions, we hope to provide a more rigorous test of the performance of the Recalibrated Meta-Probability Weighting algorithm. If we find that the Recalibrated Meta-Probability Weighting algorithm once again significantly outperforms these other algorithms, then there is strong evidence for the efficacy of this algorithm over existing aggregation approaches available in the single-question forecasting literature.

3.6.1 Methods

We collected people's responses to the questions from the Grades 1 to 5 datasets which we had used throughout the current chapter and the previous chapter (see Table 3.1). Each Grades dataset comprised 100 science statements at a US primary and secondary grade school level. These questions were adapted from worksheets on the Education Quizzes website (<http://www.educationquizzes.com/us>), and then converted into true or false statements. Approximately 2-3 questions were taken from each worksheet from the Biology, Chemistry, Geography, Physics, and General Science categories, spanning from grades 1 to 12, broken up into five levels of difficulty (grades 1 and 2; grades 3, 4, and 5; grades 6, 7, and 8; grades 9 and 10; and grades 11 and 12). We coded "Difficulty 1" as the easiest difficulty, and "Difficulty 5" as the hardest difficulty. We treated each set of 100 questions of the same difficulty as an individual dataset. The full set of

questions for this experiment are included in the Appendix (see section 6.2).

We recruited 500 respondents from Amazon Mechanical Turk and only respondents inside the US were able to participate in the experiment. Participants were paid a flat fee of \$4.00 for completing the survey. The survey was conducted on the Qualtrics platform. Before beginning the experiment, participants were first required to answer three basic logic questions to deter any non-human agents from responding to the survey. Participants were then asked to answer each question as honestly as they could and without cheating (e.g., by looking up any of the questions online). 41 individuals who reported cheating at the task or had failed to complete the survey were excluded from the analyses; analyses were conducted on the data of the remaining 459 participants.

Participants completed 100 trials each, with each trial comprising one statement which was either true or false. Half the statements at each level of difficulty were true, and the other half were false. An example of a trial can be seen in Figure 3.10. Participants were asked to provide their predictions about (1) whether the statement was more likely to be true or false, (2) what percentage of other forecasters would predict the statement to be true, (3) the probability that the statement was true, and (4) what the average probability estimated by other forecasters would be. Each participant saw 20 statements from each level of difficulty, and statements were presented in one of five randomised orders. Participants who took part in any of our previous experiments were excluded from participating.

Participants who provided votes that were inconsistent with their probability forecasts (i.e., voting “true” but predicting a probability $< 50\%$ of the statement being true, or voting “false” but predicting a probability $> 50\%$ of the statement being true) were excluded from the analysis from that particular question. 17.75% of responses were excluded in total.

3.6.2 Analyses

We conducted the same analysis as above using the new set of responses collected for this experiment. Our analyses included: (1) plots showing the overall performance for each algorithm across all five datasets; (2) computing the bootstrap 95% CIs for mean difference in Brier score between the Meta-Probability Weighting algorithm and each other algorithm, and between the

10: In physics, U-values measure how effective a material is an insulator.

Is this statement more likely to be false or true?

False



True



Everyone thinks it's false

0 10 20 30 40 50 60 70 80 90 100

Everyone thinks it's true

What percentage of other people do you think thought the bolded statement was true?



Definitely false

0 10 20 30 40 50 60 70 80 90 100

I'm not sure

Definitely true

What is the probability that the bolded statement is true?



The average probability is low

0 10 20 30 40 50 60 70 80 90 100

The average probability is high

What do you think is the average probability estimated by other people?



Figure 3.10: Example of a trial in Experiment 4 – replication of the US Grades experiment.

Recalibrated Meta-Probability Weighting algorithm and each other algorithm; (3) calibration curves for the Meta-Probability Weighting algorithm, the Recalibrated Meta-Probability Weighting algorithm, the Meta-Vote Weighting algorithm, and Recalibrated Meta-Vote Weighting algorithm; (4) plots showing the performance of each algorithm on each of the five different problem difficulties; and (5) plots showing the performance of the Recalibrated Meta-Probability Weighting algorithm compared to the optimally recalibrated versions of the other algorithms using the same cross-validation parameter estimation procedure from Experiment 2.

We fitted the Minimal Pivoting model, the Meta-Probability Weighting algorithm and the Recalibrated Meta-Probability Weighting algorithm using the appropriate meta-prediction responses provided by forecasters about the average probability forecast of others, rather than forecasters' meta-predictions about the proportion of others voting "true", which we had used until now. Each algorithm therefore now used the correct input responses, without assuming that forecasters' responses to the two meta-prediction questions were identical.

3.6.3 Results

Figure 3.11 shows the mean transformed Brier score for each algorithm across the 500 questions. In the following subsections, we report and discuss the results for each component of the analyses.

Performance of the Meta-Probability Weighting algorithm

We first compared the performance of the Meta-Probability Weighting without recalibration against the other aggregation approaches. The Meta-Probability Weighting algorithm provided significantly better forecasts than the unweighted mean by 3.43 points (95% CI: [2.73, 4.16]), the Minimal Pivoting model by 0.96 points (95% CI: [0.42, 1.52]), and the Meta-Vote Weighting algorithm by 4.03 points (95% CI: [3.58, 4.44]); not significantly different forecasts than the P''_{cs} aggregator (95% CI: [-0.43, 2.16]) and the Recalibrated Meta-Vote Weighting algorithm (95% CI: [-0.54, 0.87]); and significantly worse forecasts than the Recalibrated Meta-Probability Weighting algorithm by -3.31 points (95% CI: [-3.94, -2.59]). The Meta-Probability Weighting algorithm therefore either significantly outperformed all other non-recalibrated algorithms or at least did not provide significantly worse predictions than other algorithms overall.

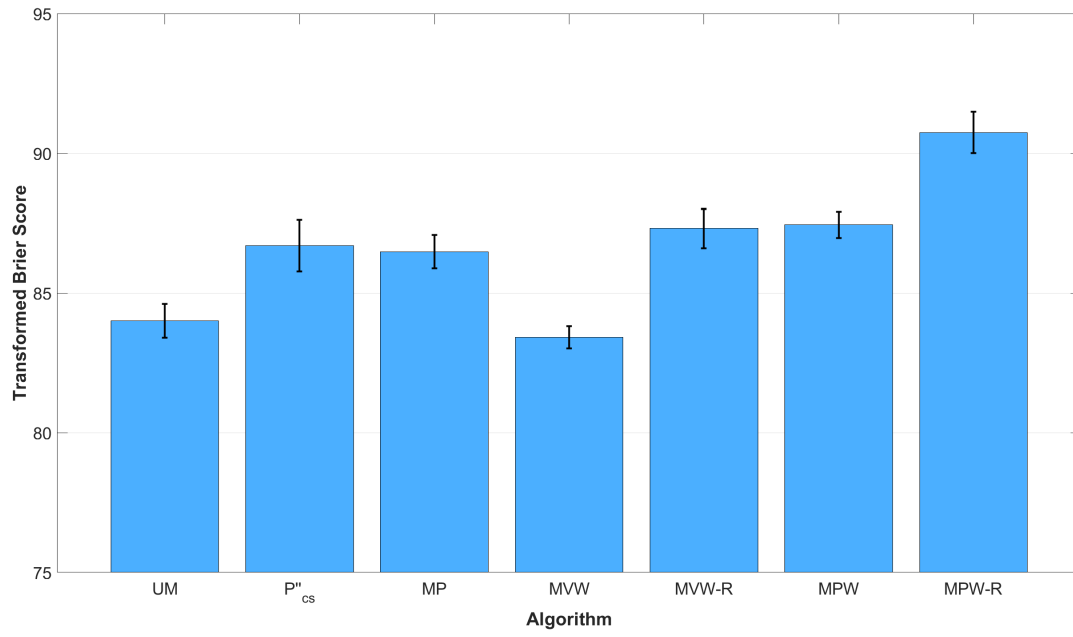


Figure 3.11: The mean transformed Brier score for the unweighted mean (UM), the P''_{cs} aggregator, the Minimal Pivoting (MP) algorithm, the Meta-Vote Weighting (MVW) algorithm, the Recalibrated Meta-Vote Weighting (MVW-R) algorithm, the Meta-Probability Weighting algorithm (MPW), and the Recalibrated Meta-Probability Weighting (MPW-R) algorithm over 500 US Grades questions varying across five levels of difficulty. Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms.

Performance of the Recalibrated Meta-Probability Weighting algorithm

We applied the recalibration function (Equation 4.3) to the Meta-Probability Weighting algorithm using a fixed parameter $a = 2.5$, and examined whether this version of the algorithm produced significantly better forecasts than all other aggregation approaches. Indeed, we found that the Recalibrated Meta-Probability Weighting algorithm provided significantly better predictions than the unweighted mean by 6.74 points (95% CI: [5.72, 7.72]), the P''_{cs} aggregator by 4.05 points (95% CI: [2.83, 5.50]), the Minimal Pivoting model by 4.27 points (95% CI: [3.44, 5.04]), the Meta-Vote Weighting algorithm by 7.33 points (95% CI: [6.29, 8.22]), and the Recalibrated Meta-Vote Weighting algorithm by 3.43 points (95% CI: [2.69, 4.18]). The Recalibrated Meta-Probability Weighting algorithm therefore outperformed all other competing algorithms by a larger amount than we had seen in the results from Experiment 3.

Calibration curves

The calibration curves for the Meta-Vote Weighting algorithm, Recalibrated Meta-Vote Weighting algorithm, Meta-Probability Weighting algorithm, and Recalibrated Meta-Probability Weighting algorithm are shown in Figure 3.12. Comparing the curves for the Meta-Vote Weighting algorithm (orange) and Meta-Probability Weighting algorithm (light green), we can see that the Meta-Probability Weighting algorithm is slightly better calibrated for the 30–40, 40–50, 50–60, 60–70, and 70–80 percentile bins. Comparing the curves for the Recalibrated Meta-Vote Weighting algorithm (red) and the Recalibrated Meta-Probability Weighting algorithm (blue), we can see that the Recalibrated Meta-Probability Weighting algorithm is much better calibrated for almost every percentile bin.

Performance at the dataset level

The performance of each algorithm on each level of difficulty is shown in Figure 3.13. The Recalibrated Meta-Probability Weighting algorithm (purple bar) was the best-performing algorithm on all five levels of difficulty, and offered a significant improvement over all other algorithms for difficulties 2, 3, and 5. On difficulties 1 and 4, the Recalibrated Meta-Probability Weighting algorithm still outperformed all other algorithms, but these differences were not statistically

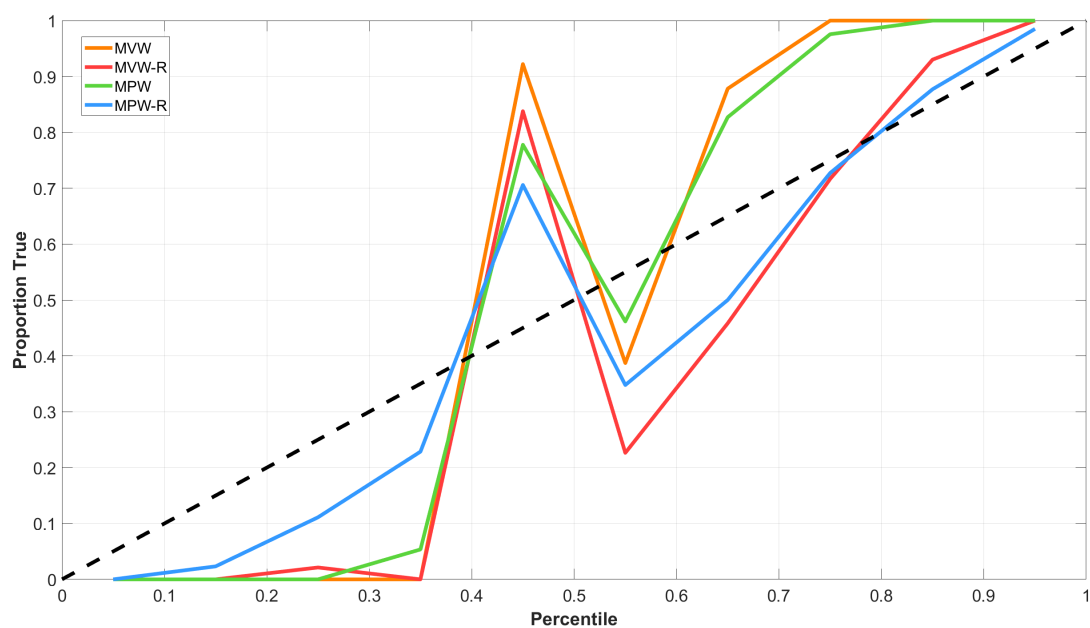


Figure 3.12: The calibration curve for the Meta-Vote Weighting algorithm (orange), the Recalibrated Meta-Vote Weighting algorithm (red), the Meta-Probability Weighting algorithm (light green), and the Recalibrated Meta-Probability Weighting algorithm (blue). Each curve shows the proportion of events where the outcome was “true” vs. the proportion of events forecasted to be “true” by that algorithm. The dotted reference line indicates the calibration of a perfectly-calibrated forecaster (or algorithm), whose forecasted probabilities match exactly the proportion of “true” events observed.

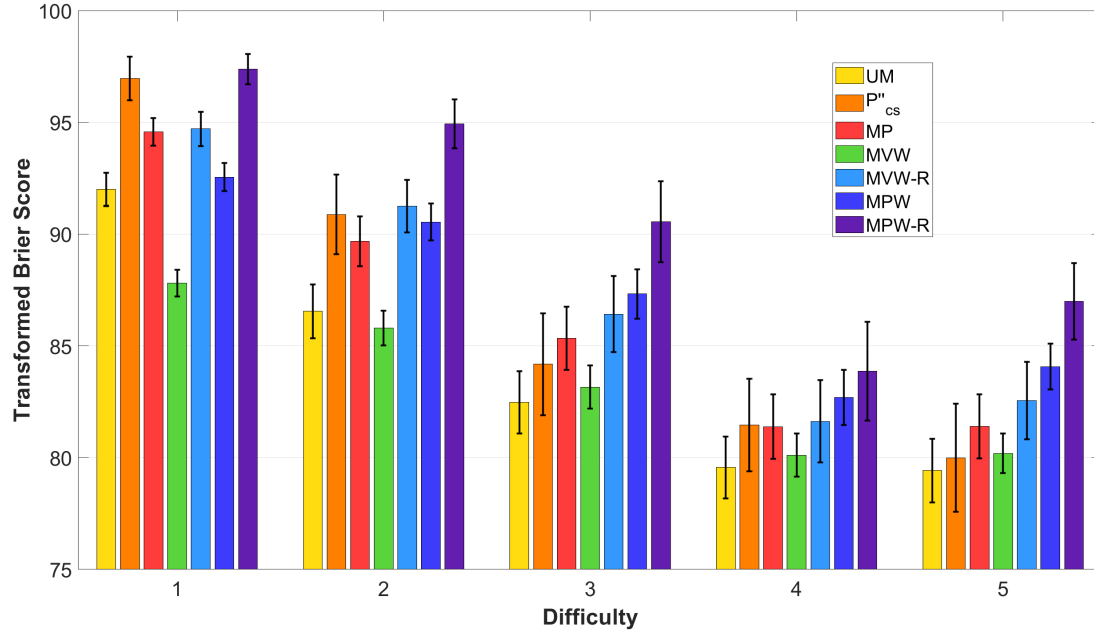


Figure 3.13: The mean transformed Brier score for the unweighted mean (yellow), the P''_{cs} aggregator (orange), the Minimal Pivoting model (red), the Meta-Vote Weighting algorithm (light green), the Recalibrated Meta-Vote Weighting algorithm (light blue), the Meta-Probability Weighting algorithm (dark blue), and the Recalibrated Meta-Probability Weighting algorithm (purple) on each US Grades dataset, in order of increasing difficulty from lowest difficulty level (Difficulty 1) to highest difficulty level (Difficulty 5). Error bars show the standard error.

significant.

Optimally recalibrating model predictions

Once again, to rule out the explanation that the parameter of the recalibration function was simply better-suited to the Meta-Probability Weighting algorithm than other algorithms, we used leave-one-out cross-validation to estimate the optimal recalibration parameters for each other algorithm, in the same way as in Experiment 2 (see Section 3.4.3 for details).

Figure 3.14 shows the mean score for each of the Recalibrated unweighted mean, Recalibrated P''_{cs} aggregator, Recalibrated Minimal Pivoting model, and Recalibrated Meta-Probability Weighting algorithm across all 500 questions. Despite optimal recalibration, the unweighted mean, P''_{cs} aggregator, and Minimal Pivoting model still performed significantly worse than the non-optimised Recalibrated Meta-Probability Weighting algorithm. Specifically, the Recalibrated

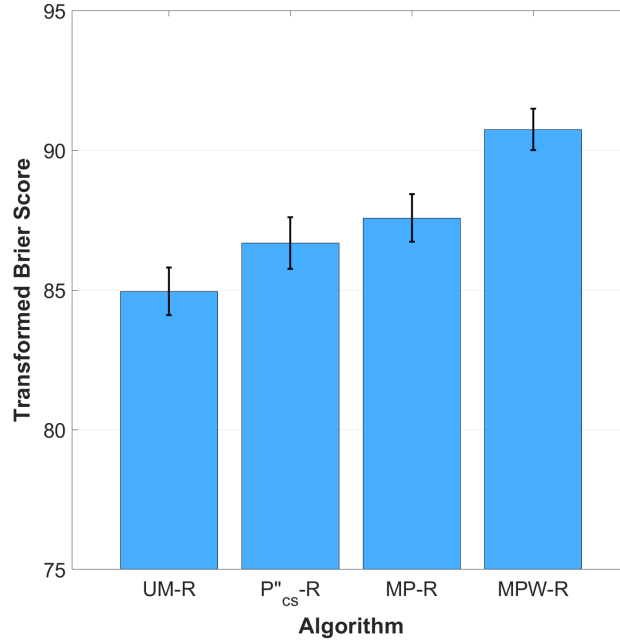


Figure 3.14: The mean transformed Brier score for the optimally recalibrated versions of the unweighted mean (UM), the P''_{cs} aggregator, and the Minimal Pivoting (MP) algorithm, compared to the Meta-Probability Weighting (MPW) algorithm, which has not been optimally recalibrated. Error bars show the standard error. The Recalibrated Meta-Probability Weighting algorithm (far right) significantly outperforms all other algorithms.

Meta-Probability Weighting algorithm significantly outperformed the optimally recalibrated unweighted mean by 5.79 points (95% CI: [4.66, 6.93]), the optimally recalibrated P''_{cs} aggregator by 4.06 points (95% CI: [2.84, 5.51]), and the optimally recalibrated Minimal Pivoting model by 3.17 points (95% CI: [2.84, 5.51]). These results show that even in the most extreme case, these algorithms cannot be recalibrated to outperform the Recalibrated Meta-Probability Weighting algorithm.

3.6.4 Discussion

The aim of the current experiment was to test whether the Recalibrated Meta-Probability Weighting algorithm would outperform other standard forecast-aggregation approaches when these algorithms were provided with the appropriate meta-prediction responses from forecasters. Our results showed that the Recalibrated Meta-Probability Weighting algorithm indeed improved

in performance when forecasters’ meta-predictions about the average probability forecast of others was used, rather than their meta-predictions about the proportion of others voting “true”. This improvement in score could be seen in the large difference in performance between the Recalibrated Meta-Probability Weighting algorithm and each other aggregation approach, recalibrated or otherwise.

3.6.5 Supplementary analyses

Having established the efficacy of the Recalibrated Meta-Probability Weighting algorithm, we conducted three additional sets of analyses investigating: (1) the robustness of these algorithms over smaller sample sizes, (2) the weights assigned by the Meta-Probability Weighting algorithm to each quartile of forecasters in the dataset, and (3) the performance of smaller ‘select’ crowds, chosen according to the weights assigned by the Meta-Probability Weighting algorithm to forecasters. These analyses provide further insight the properties of the weights used by the SP algorithm.

Robustness over smaller sample sizes

We investigated the robustness of the Meta-Probability Weighting algorithm and Recalibrated Meta-Probability Weighting algorithm over different sample sizes. In particular, we wanted to identify whether these approaches maintained their advantage over other algorithms for smaller-sized crowds. Simulations over different sample sizes in Chapter 2 showed that the SP algorithm was robust over small samples and outperformed the majority vote for samples as small as 20 people. Since the Meta-Probability Weighting algorithm operates under similar theoretical principles, we would expect similar robustness in performance for smaller samples.

We simulated changes in performance for each algorithm over different sample sizes using bootstrap resampling. For each sample size n from 5 to 100 forecasters in increments of 5, we randomly resampled n forecasters from the original sample, with replacement. On each of these bootstrap samples, we aggregated forecasters’ responses using the unweighted mean, Meta-Vote Weighting algorithm, Recalibrated Meta-Vote Weighting algorithm, Meta-Probability Weighting algorithm, and Recalibrated Meta-Probability Weighting algorithm. We repeated this 1,000 times for each question and each n , which we then averaged to obtain a score for each algorithm. We

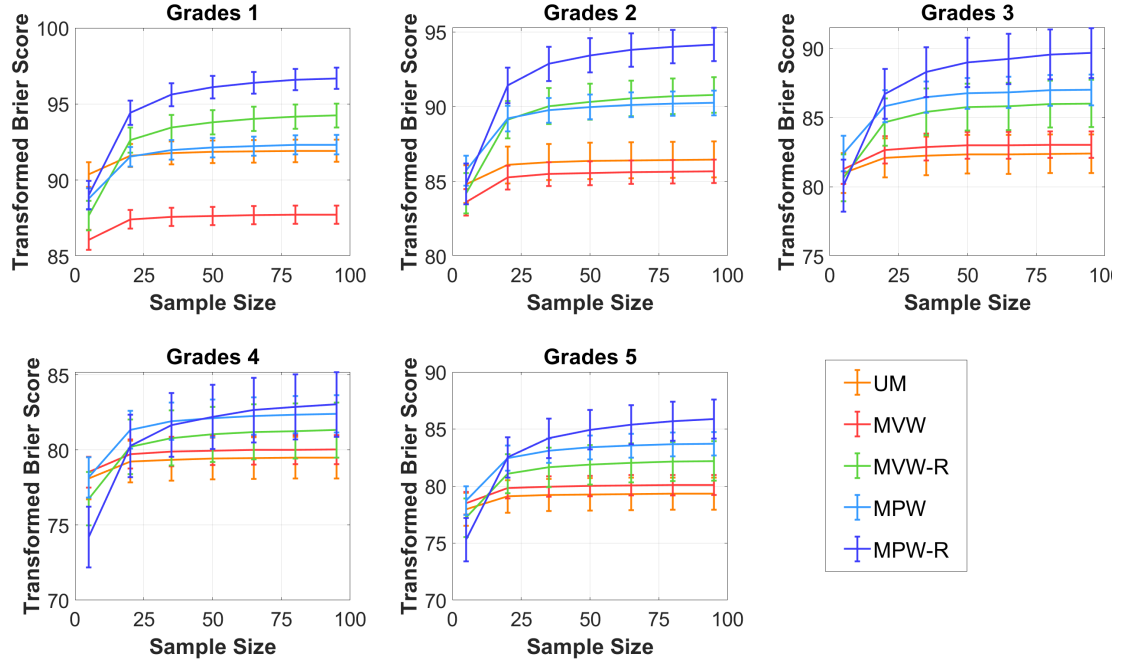


Figure 3.15: Simulation results showing the change in score for the unweighted mean (UM), Meta-Vote Weighting (MVW) algorithm, Recalibrated Meta-Vote Weighting (MVW-R) algorithm, Meta-Probability Weighting (MPW) algorithm, and Recalibrated Meta-Probability Weighting (MPW-R) algorithm over different sample sizes for each level of question difficulty in Experiment 4. Error bars show the standard error.

then averaged the score over all 100 questions at each difficulty for each algorithm, and plotted the change in performance as a function of sample size.

Figure 3.15 shows the change in each algorithm’s performance on each of the five difficulties. The recalibrated Meta-Probability Weighting algorithm performed well relative to the unweighted mean on all five datasets for moderate to large sample sizes. However, when samples contain less than approximately 30 forecasters, the Recalibrated Meta-Probability Weighting algorithm appears to exhibit a large drop in performance for all five datasets. The Meta-Probability Weighting algorithm also shows a similar pattern, indicating that Meta-Probability Weighting metric may be an unreliable approach under small sample sizes. These results are fairly consistent across all five difficulties.

A comparison between the Meta-Probability Weighting algorithm and the Decision Similarity metric

We compared the weights used by the Meta-Probability Weighting algorithm, the Meta-Vote Weighting (and SP) algorithm, relative to weights derived from the Decision Similarity metric proposed by Kurvers et al. (2019). The Decision Similarity metric was developed to distinguish between high-performing and low-performing individuals without the need for records of forecasters' past performance. The comparison between these two metrics provides a powerful test of the efficacy of the Meta-Probability Weighting metric, which has also been developed to distinguish between high-performing and low-performing individuals without requiring records of their past performance. As a formal model for weighting forecasters by Decision Similarity remains to be developed, we implemented the Decision Similarity metric as normalised weights in the same way that forecasters are weighted by the Meta-Probability Weighting algorithm and the Meta-Vote Weighting algorithm. Applying forecasters' Decision Similarity provides at least some basic insight into the relative efficacy between the Decision Similarity and Meta-Probability Weighting metric for identifying crowd expertise in the single-question domain.

We examined the weights derived from forecasters' Decision Similarity and compared them to the weights from the Meta-Probability Weighting algorithm and the Meta-Vote Weighting algorithm to see which weighting metric better corresponded to forecasters' percentage accuracy. To achieve this, we conducted the same analyses from Chapter 2 in which we sorted forecasters according to their percentage accuracy and plotted the average weight assigned to each quartile of forecasters. Forecasters' percentage accuracy was estimated separately for each difficulty using leave-one-out cross-validation (for more details, see Section 2.4.4).

Figure 3.16 shows the average weight assigned to forecasters in each quartile for each of the three weighting metrics. The top quartile of forecasters (Q4) were assigned much higher Meta-Probability weights than Decision Similarity or Meta-Vote weights in every dataset except the easiest difficulty. Meta-Probability weights were therefore generally much more effective for distinguishing between high-performing and low-performing individuals than these other metrics.

In contrast, Decision Similarity could only be used to distinguish between high-performing and low-performing individual on the two easiest difficulties. On the three hardest difficulties,

Decision Similarity was highly ineffective, and could not differentiate between forecasters of each quartile in the correct manner, indicated by the flat or decreasing slope of the line as forecasters' expertise increased. While the exact magnitudes of these weights may be small due to the way we chose to apply these weights, the relative ordering of the weights for the Decision Similarity metric for each of the quartiles would be maintained under any linear transformation of the weights. That is, even if the weights derived from the Decision Similarity metric were transformed (e.g., using a power function) to increase up-weighting of individuals with the highest Decision Similarity, this would not change the fact that the individuals with high Decision Similarity are under-performing individuals with low Decision Similarity for all but the lowest difficulties. In the next section, we provide another test of the Decision Similarity metric using a method that does not treat Decision Similarity as normalised weights, thereby ruling out any explanation that the poor performance of the Decision Similarity metric may simply be due to the way that Decision Similarity has been applied as forecasters' weights.

Cross-validating the Meta-Probability Weights

In this section, we examine a stronger test for the efficacy of each weighting metric at identifying expertise and distinguishing between high-performing individuals and low-performing individuals. We sorted forecasters by each of the three metrics – Meta-Probability Weighting, Meta-Vote Weighting, and Decision Similarity – and examined the performance of the forecasts from a crowd comprising the top-performing forecasters over a range of crowd sizes. If the Meta-Probability Weighting metric is more effective than these other measures at quantifying expertise, then we should see a crowd comprising the top Meta-Probability-weighted forecasters to outperform a crowd comprising the same number of top Decision Similarity forecasters (or Meta-Vote-weighted forecasters).

To test which of these weighting metrics was most effective for predicting high-performance individuals, we conducted a leave-one-out cross-validation (i.e., a jack-knife) procedure using training and test sets. There were 500 iterations in the cross-validation procedure – one iteration for each of the 500 questions in Experiment 4's dataset. Each event in the dataset was assigned as a test set for one iteration, and the remaining events were assigned to the training set. In the training set, we calculated the average Meta-Probability Weight, the average Meta-Vote Weight,

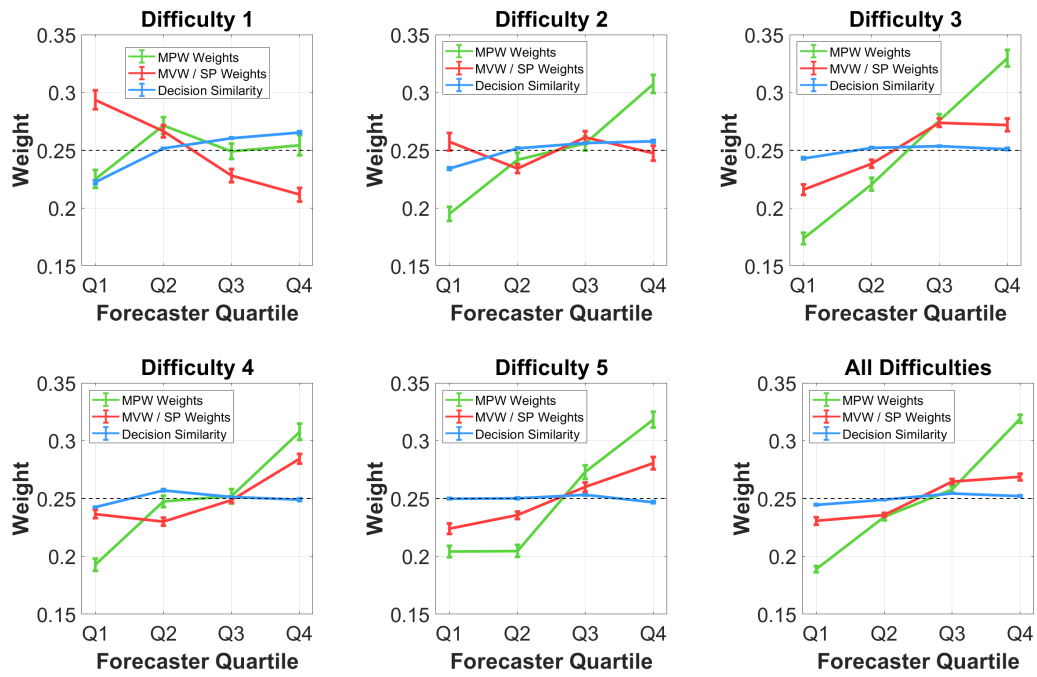


Figure 3.16: The average weight assigned by the Meta-Vote Weighting (and SP) algorithm, the Meta-Probability Weighting algorithm, and the Decision Similarity weighting model to the votes of forecasters as a function of the performance of the forecasters for each of the five difficulties and overall across all five difficulties. Forecasters were ranked by performance in terms of percentage accuracy on each event. Error bars represent the standard error.

and the average Decision Similarity for each individual across all questions in the training set, and selected the top n percentage of forecasters in the range of $n \in \{100, 75, 50, 25, 10\}$. We then looked at the performance on the test set for the aggregate forecast based on each weighting metric, which we generated by taking an unweighted mean of the forecasts from all the forecasters selected from the training set. We then plotted the average performance for each weighting metric over n .

Figure 3.17 shows the performance of each weighting metric in each of the five difficulties and generally across all difficulties. Overall, selecting the top forecasters according to their Meta-Probability weights (green) consistently resulted in more accurate forecasts than selecting the top forecasters based on Meta-Vote weights (red) and Decision Similarity (blue). We can see that this trend is robust for almost every size sub-crowd containing different percentages of the top forecasters. Furthermore, this was also the case in each of the five problem difficulties, including the easiest difficulty, where Decision Similarity should, in theory, be most effective for identifying high-performing individuals. The Meta-Probability Weighting metric appears to provide the most effective metric for identifying expertise in the single-question domain in the current literature.

3.7 General Discussion

The primary aims of this chapter were to develop a measure by which forecasters' expertise can be robustly quantified and thus leveraged to produce accurate aggregated forecasts, to provide an empirical comparison between these models for aggregation in the single-question domain, and to present potential avenues for extending the findings from Chapter 2 to the probabilistic domain. Experiment 1 tested whether single-question categorical forecasting algorithms such as the SP algorithm (Prelec et al., 2017) could be applied effectively to probabilistic forecasting and found that even though the SP algorithm performed well on categorical measures of performance, it scored poorly relative to other probabilistic algorithms on probabilistic problems on measures such as the Brier score.

We subsequently developed the Meta-Vote Weighting algorithm, which weights forecasters in the same way as the SP algorithm, using the normalised absolute difference between forecasters'

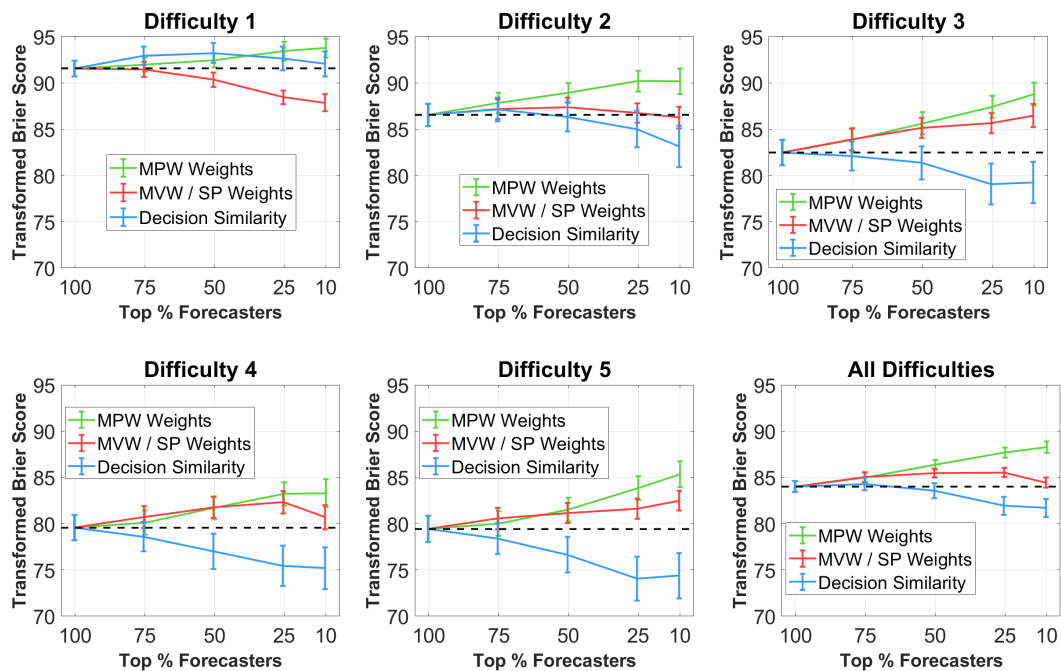


Figure 3.17: Performance for each of the weighting metrics after averaging over the forecasts of the top n percentage of forecasters selected from the training set using leave-one-out cross-validation. Forecasters were selected based on their average weight in the training set, determined by either their Meta-Probability weights, Meta-Vote weights, or their Decision Similarity. Error bars represent the standard error. Meta-Probability Weights appear to consistently outperform both other weighting metrics.

votes and meta-predictions about the proportion of others voting “true”. We tested the performance of this novel algorithm against other standard probabilistic forecasting algorithms in Experiment 2 and found that it generally did not perform better than these other algorithms. Although the Meta-Vote Weighting algorithm was able to identify experts in the crowd, the aggregated forecast from experts was still systematically under-confident, and needed to be extremised. By combining the Meta-Vote Weighting algorithm with a basic recalibration function adapted from Baron et al. (2014), we were able to generate significantly better probability forecasts than any other existing single-question aggregation approaches. We showed that even when these other aggregation approaches were recalibrated using the optimal parameters tailored to each algorithm, the Recalibrated Meta-Vote Weighting algorithm still generated significantly better forecasts than these other recalibrated algorithms. These results were robust over a range of moderate-to-large-sized samples.

In Experiment 3, we developed the Meta-Probability Weighting algorithm, which improved the weighting mechanism used by the Meta-Vote Weighting algorithm. The Meta-Probability Weighting algorithm weights forecasters’ probability forecasts by the normalised absolute difference between forecasters’ probability forecasts and their meta-predictions about the average probability forecast of others. This improved approach provided more accurate probability forecasts than even the recalibrated Meta-Vote Weighting algorithm. By extremising the forecasts of this algorithm, we obtained a larger improvement over other algorithms, even when those other algorithms were optimally recalibrated. Comparing the performance of the Meta-Probability Weighting algorithm on individual datasets, we found that it was highly robust across a large range of questions varying on different levels of difficulty. In Experiment 4, we replicated these findings using the five US Grade School question sets with a new sample of forecasters and found that these results replicated reliably.

After examining the efficacy of the weights used by the Meta-Probability Weighting algorithm, we found that they were empirically superior to both the weights used by the Meta-Vote Weighting algorithm and weights derived from Decision Similarity (Kurvers et al., 2019). The weights used by the Meta-Probability Weighting algorithm therefore provided the most effective metric for identifying expertise in the single-question domain compared to other existing metrics in the literature.

3.7.1 Contributions of the present research

The current chapter provides several major contributions to the forecasting literature. We have developed a Meta-Probability Weighting algorithm that generally outperforms all existing algorithms in the current single-question forecasting literature. While the P''_{cs} aggregator and Minimal Pivoting model also provided improvements over the unweighted mean on particular datasets, those improvements were less consistent and typically smaller in magnitude than the improvements provided by the Recalibrated Meta-Probability Weighting algorithm. Our results therefore have important practical implications for researchers and decision analysts who are seeking to generate the most accurate probabilistic forecasts in the single-question domain.

We have also developed a novel measure for identifying forecaster expertise in the single-domain, where other measures of expertise may be ineffective or unavailable. Our work in this chapter builds on that of the previous chapter, where we showed that the absolute difference between forecasters' votes and meta-predictions about the proportion of others voting "true" was predictive of expertise in certain forecasting environments. Here, we have shown that our novel measure of expertise, the absolute difference between forecasters' probability forecasts and their meta-predictions about the average probability forecasted by others, is much more effective at quantifying expertise than existing measures of expertise and is also more robust across a wider range of forecasting environments.

In Experiments 3 and 4, we showed that this measure of expertise can be used to leverage forecasters' probabilistic predictions in order to produce accurate aggregated forecasts. We have demonstrated one usage here for this novel measure of expertise, and there are sure to be many other applications to be considered. For example, the goal of decision analysts may not only to generate the most accurate prediction for any given problem, but to identify experts for other reasons such as for training or compensation. While other contemporary aggregation approaches such as the P''_{cs} aggregator or Minimal Pivoting model are able to provide accurate probability forecasts, they have not been designed to identify forecasters' expertise. Similarly, the Decision Similarity metric appears to be only effective for identifying high-performing individuals on low-difficulty problems. The Meta-Probability Weighting algorithm therefore not only generates better probabilistic forecasts than these other algorithms, but can also be used to identify expertise

in a way that is not possible under existing single-question forecasting approaches.

We have provided a comprehensive large-scale comparison of the existing single-question forecast-aggregation approaches in the literature. To date, no other datasets available in the literature has collected both forecasters meta-predictions about the proportion of other forecasters voting “true” and the average probability forecasted by others. Our experimental data is therefore uniquely valuable in allowing for these aggregation approaches to be compared using the appropriate responses from forecasters. Our results from this experimental data provide novel insight into the performance of these algorithms when each algorithm is given a fair comparison, which was not possible using the existing datasets available from other studies.

Lastly, our results throughout this chapter also provide valuable evidence of the efficacy of these other algorithms in a wide variety of probabilistic forecasting problems. Other than the initial studies in which these algorithms were proposed, few studies to date have compared or applied these algorithms. Our results have demonstrated that these other algorithms can also be highly effective at generating probability forecasts in many types of forecasting settings, although not to the same extent as the Meta-Probability Weighting or Meta-Vote Weighting algorithm. Our research therefore provides some merit in demonstrating the general replicability of this previous research, providing a stepping stone toward developing novel and refined probabilistic forecasting approaches, particularly in the single-question domain.

3.7.2 Relationship to previous research

The Meta-Probability Weighting algorithm operates on the same fundamental principles as other weighting approaches in the forecasting literature (e.g., Budescu & Chen, 2015; Clemen, 1989; Davis-Stober et al., 2014; Winkler, 1989). Weighting approaches to forecast aggregation seek to maximise forecast performance by identifying experts in the crowd and combining their forecasts in a theoretically ‘optimal’ manner. A major difference between the Meta-Probability Weighting algorithm and existing algorithms is that the Meta-Probability Weighting algorithm has been designed to be able to identify expertise in the single-question domain, whereas other expertise-identification approaches typically require forecasters’ responses on multiple questions with known outcomes. Other algorithms that have been developed in the single-question domain also seem to

operate under different principles than these expertise-identification approaches, for example, by identifying the overlap in information between forecasters (Palley & Soll, 2019; Satopää et al., 2016). The Meta-Probability Weighting algorithm thus builds upon the single-question forecasting literature in a novel way, since existing measures of expertise – such as Decision Similarity – are only effective under much more narrow conditions compared to the Meta-Probability Weighting algorithm. Our findings provide an important contribution in connecting the expertise-driven approaches in the existing literature with expertise-driven approaches in the single-question domain.

The finding that recalibration significantly improves the performance of the Meta-Probability Weighting algorithm is consistent with the wider literature, which demonstrates the need of extremisation for aggregated forecasts (e.g., Baron et al., 2014; Ranjan & Gneiting, 2010). While we found that the unweighted mean and Minimal Pivoting models also benefited from extremisation, the Meta-Probability Weighting algorithm received a larger improvement in score from extremisation than these other algorithms. For example, the P''_{cs} aggregator and Minimal Pivoting model both received only minute improvements in score, even when optimally recalibrated. Extremisation thus appears to be important and complementary to the Meta-Probability Weighting algorithm – more so than for other algorithms.

3.7.3 Considerations for future research

One consideration for future studies would be to compare the performance of these algorithms using different recalibration functions. While we believe this would not change the general outcome of our results, there may be potential recalibration functions better suited to maximising the performance of these algorithms in the single-question domain. Here, we have chosen a recalibration function that has been applied extensively in past research and has previously been optimised over a range of forecasting environments (Baron et al., 2014). Future researchers may consider testing the efficacy of other recalibration approaches whose parameters can be estimated without the need for forecasters’ responses to other questions with known outcomes.

Another important consideration for future research is the application of forecast-aggregation approaches to novel forecasting problems. While we have developed effective solutions to these

problems in the single-question domain, it may be possible to address the same fundamental problems with leveraging forecasters' expertise in other ways. For example, even if decision analysts do not have records of forecasters' performance on questions related to the question of interest, they may be able to obtain forecasters' responses to a set of other seemingly-unrelated questions that have known outcomes. The current literature provides little guidance on whether forecasters' expertise can be measured reliably by their performance on sets of unrelated questions. A useful extension to the current research would therefore be to investigate whether these algorithms developed for the single-question domain may be applicable under more relaxed constraints. The following chapter investigates whether accurate probabilistic forecasts can be generated in the multi-question domain when only forecasters' responses to other sets of unrelated questions are available. In the process, we examine whether the Meta-Probability Weighting algorithm provides any merit when applied beyond the single-question domain.

4 | Identifying expertise via cross-domain performance

In the current chapter, we examine the extent to which forecasters' expertise can be identified via their performance on seemingly unrelated domains. While the previous two chapters applied forecast-aggregation algorithms in the single-question context, the current chapter moves into the multiple-question domain, where forecasters provide responses to multiple questions where the outcomes to some of these questions may be known to the decision maker.

4.1 Abstract

This chapter explores probabilistic forecasting in environments where decision makers have access to forecasters' responses to a set of questions that are unrelated to the question of interest. Past research have found the existence of superforecasters – forecasters who consistently outperform others in the quality of their predictions and are able to generate predictions across multiple related domains such as geopolitical and economic forecasting (Mellers et al., 2015). This chapter explores the concept of cross-domain forecasting across a wider range of unrelated domains. We aim to explore whether experts can be identified generally across multiple different domains including art, science, sport, and emotional intelligence. Over four experiments, we found that expertise could be found consistently across domains, and weighting forecasters by their contributions to the crowd prediction in other domains was generally as effective as weighting using contributions from the same domain. We present cross-domain weighting as a viable

alternative to expertise-identification algorithms in the single-question context, where no historic information on forecasters’ performance on related questions is available. Our results highlight the benefits of cross-domain weighting over single-question aggregation algorithms, and quantify the trade-off between forecasting performance and the option to elicit forecasters’ responses to questions with known outcomes.

4.2 Introduction

In forecasting and decision problems, decision makers often have access to responses from multiple different individuals and are faced with the challenge of aggregating those responses into a single decision. In many cases, the decision maker does not have any records of forecasters’ past performance in the domain of interest, for example, because the forecasting problem is entirely novel to that set of forecasters or because there is no set of relevant questions for which the outcomes are known. These practical constraints pose a challenge for many traditional aggregation approaches, which typically rely on forecasters’ past performance to select and up-weight high-performing individuals in the crowd (Armstrong, 2001; Budescu & Chen, 2015; Clemen, 1989; Cooke, 1991; Winkler, 1989).

In recent years, a number of approaches have been developed to aggregate forecasts in what has become known as the ‘single-question’ context (Kurvers et al., 2019; J. McCoy & Prelec, 2017; Palley & Soll, 2019; Prelec et al., 2017; Satopää et al., 2016). Prelec et al. (2017) developed an insightful approach for aggregating binary predictions when the majority of forecasters are biased. The Surprisingly Popular (SP) algorithm they developed is theoretically robust to bias and it was found to perform well in practice across a variety of domains including the classification of medical images, judgments about the price of artworks, and science trivia problems.

More recently, Kurvers et al. (2019) proposed the *Decision Similarity* measure, which quantifies the similarity of forecasters’ binary decisions to the decisions of other forecasters in the crowd. The authors demonstrated, both theoretically and empirically, that the Decision Similarity was an effective measure for distinguishing between high-performing and low-performing individuals in the crowd. The authors also showed empirically that when forecasters are more often correct than incorrect, crowds containing the individuals with the highest decision similarity consistently

outperformed majority voting. In contrast, they found that when majority voting performed poorly, decision similarity was also not an effective measure for distinguishing between high-performing and low-performing individuals. Because it is often hard to establish in advance whether the majority of forecasters will be correct or incorrect, particularly in high-stakes problems, the practical utility of the Decision Similarity approach may be limited.

Single-question approaches also have been developed for aggregating continuous judgments, such as probabilistic forecasts. Palley and Soll (2019) developed a theoretical framework where they model the structure of information available to forecasters in the crowd. Their framework uses the idea that the simple average will perform sub-optimally when the information available to forecasters is shared – for example, this could be the case if forecasters access the same media or read the same articles. Multiple forecasters may then base their forecasts on the same information, resulting in over-representation of that information, while unique information that is available to only a small subset of forecasters will result in under-representation of that information. Under this framework, the authors develop a series of algorithms that correct for this bias. The optimal algorithm for combining forecasts when information is shared depends heavily on the structure of information shared by forecasters in the crowd. Since the exact structure of information available to forecasters might be unknown to the decision maker seeking to aggregate these forecasts, the authors proposed a reasonable solution to be the Minimal Pivoting approach, which provides the most conservative correction relative to other pivoting procedures, all of which require knowledge of the structure of information shared by forecasters.

More recently, Martinie et al. (2020) developed the Meta-Probability Weighting (MPW) algorithm as a method of leveraging the latent crowd expertise using forecasters’ meta-predictions about the forecasts of others. The authors found that their algorithm improved upon the performance of existing single-question aggregation algorithms, including the Minimal Pivoting approach, across a range of decision problems varying in difficulty. The authors showed that under reasonable theoretical assumptions the MPW algorithm will always assign greater weights to experts than novices in the crowd and is therefore theoretically guaranteed to outperform simple averaging in a large class of decision problems.

While these single-question aggregation approaches generally perform well relative to simpler aggregation approaches such as majority voting and simple averaging, their main appeal is that

they can be applied in the case where records of forecasters' past performance in a related domain are unavailable. To date, few comparisons have been made between single-question approaches and aggregation approaches that use forecasters' past performance to identify expertise, and thus, it remains unclear whether these single-question approaches will outperform these other, more traditional approaches when it is possible to obtain records of forecasters' past performance. The first aim of the current chapter is therefore to examine the efficacy of single-question approaches relative to weighting by forecasters' performance on 'seed' questions with known outcomes.

As part of this chapter's overarching aim to evaluate the efficacy of these single-question algorithms relative to other aggregation approaches to the literature, we will also examine the performance of a cognitive model developed for probabilistic forecasting. Lee and Danileiko (2014) developed a cognitive model that jointly estimates forecasters' expertise and a probabilistic forecast for each event in a set of forecasting problems. In their model, which we will refer to as the *LD cognitive model*, forecasters' predictions are drawn from a Gaussian distribution centered on the true probability and the expertise of each forecaster is captured by the variance in the samples they draw from the Gaussian distribution. Expert forecasters can be identified via greater precision in their knowledge of the true probability and therefore lower variance in their draws. In their model, forecasters are also expected to differ in their level of calibration, with more calibrated forecasters reporting probability forecasts that better match the true probability, whereas poorly calibrated forecasters report probabilities that are systematically under-confident relative to the true probability.¹

An advantage of Lee & Danileiko's (2014) cognitive modeling approach is that it does not require the true question outcomes for any of the training problems to be known to the decision maker. The cognitive modeling approach is therefore appealing in problems where decision makers are unable to infer expertise based on forecasters' past performance or other characteristics. The cognitive modeling approach is potentially less practical than single-question approaches, since it requires forecasters' predictions to multiple questions.² Nonetheless, as no empirical comparison

¹The parameter in the calibration function estimated by the model take a linear-in-log-odds functional form with a single parameter capturing the magnitude of over- and under-estimation. This calibration function takes a very similar functional form to the recalibration used in the previous chapter from Baron et al. (2014)

²Note that because this modeling approach requires forecasters' responses on multiple questions, it was not be applied in the earlier chapters. The focus of those chapters was on single-question forecasting problems, where forecasters are only providing responses to a single problem of interest, and thus, this modeling approach could not be applied.

between this cognitive modeling approach and single-question aggregation approaches have been made to date, it remains an empirical question as to how such a cognitive model will perform relative to single-question approaches, and relative to more costly approaches, such as those that estimate forecasters' expertise based on their past performance. The current chapter will therefore examine the performance of the cognitive modeling approach from Lee and Danileiko (2014) and evaluate it in comparison to these other approaches.

One concern with obtaining records of forecasters' past performance is that decision makers might find it difficult to determine the appropriate seed questions for identifying forecasters' expertise on the problem of interest. Unfortunately, the current literature provides little guidance on what makes a good set of seed questions (Bolger & Rowe, 2015; Genest & McConway, 1990). Presumably, seed questions should be as similar as possible to the test questions. However, existing studies almost exclusively use seed questions that are extremely similar or identical to the domain of the test questions (e.g., see Budescu & Chen, 2015; Cooke, 1991; Mannes et al., 2014). It is therefore unclear to what extent decision makers can use seed questions from unrelated domains to estimate forecasters' expertise. The second aim of the current chapter is to explore the extent to which performance in one domain can be used to weight forecasts in an unrelated domain.

Some recent findings about superforecasters – a set of experts found in the Intelligence Advanced Research Projects Activity (IARPA) geopolitical forecasting tournament who consistently outperform other forecasters year after year (Mellers et al., 2015; Tetlock & Gardner, 2016) – suggest that it may be possible to identify forecasters' expertise across multiple seemingly unrelated domains. Mellers et al. (2015) propose four reasons why superforecasters consistently outperform others: (1) cognitive abilities and styles – superforecasters have higher intelligence than non-superforecasters, (2) task-specific skills – superforecasters understand the specific requirements of each task, (3) motivation and commitment – superforecasters were more committed to the task, and (4) enriched environments – superforecasters sought out and engaged others' opinions more frequently. While we would not necessarily expect these factors to determine forecasters' performance in other question domains, these factors should still determine forecasters' performance to some extent. For example, motivation and commitment may play a role in how all forecasters perform, regardless of domain. Similarly, forecasters with greater general intelligence

are likely to naturally perform better than other forecasters regardless of domain. Thus, if forecasters' performance in these domains were mostly determined by motivation and general intelligence rather than task-specific knowledge, it should be possible to obtain an accurate estimate of forecasters' expertise on the domain of interest regardless of how unrelated the seed domain may be.

On the other hand, domain-specific expertise is also driven by forecasters' past experience and background knowledge. It may be ineffective to estimate forecasters' expertise across domains if forecasters only have specialised knowledge for some of these domains, but not others. Whether forecasters' expertise can be estimated via their performance on unrelated domains is therefore likely to depend heavily on the relationship between the test and seed domains. For example, domains that draw on different sets of task-specific skills would likely make it difficult to obtain accurate estimates of expertise across those domains. On the other hand, if each domain drew on the same sets of skills, then we would likely expect little loss in accurately estimating forecasters' expertise. By comparing forecasters' performance across a range of domains that vary in similarity, we may be able to identify the extent to which cross-domain expertise identification is feasible.

In the following section, we provide a detailed explanation of our research design, explain how we will address our research aims, and provide an outline of each of our experiments.

4.3 Research Design

We conducted four experiments comparing the efficacy of weighting forecasters by their expertise on questions in the same domain versus other, unrelated domains. We applied a similar methodology in all four experiments, with the main difference being the domains of questions used. We adopted the Contribution-weighted Model (CWM; Budescu & Chen, 2015) as a general measure of forecaster expertise, since past studies have shown its efficacy over other measures such as past performance (Budescu & Chen, 2015; Chen, Budescu, Lakshmikanth, Mellers, & Tetlock, 2016). In the CWM, forecasters are weighted by the contribution metric, computed from the improvement offered by individual forecasts over the average crowd forecast. Forecasters who improve the aggregate forecast thus have positive contributions whereas forecasters who worsen the aggregate forecast have negative contributions, and are excluded from the aggregate forecast.

Formally, the contribution of the j th forecaster for a given question (where N_j is the total number of forecasters for that question) is proportional to the difference between the transformed Brier score of the crowd’s average forecast, S_i , and the transformed Brier score of the crowd’s average forecast without that forecaster, S_i^{-j} :

$$C_j = \sum_{i=1}^{N_j} \frac{S_i - S_i^{-j}}{N_j} \quad (4.1)$$

The transformed Brier score, used to assess the accuracy of the forecast, is given by:

$$S = 100 - 50 \sum_{k=1}^K \frac{(D(o_k) - T(X_k))^2}{K}, \quad (4.2)$$

where $D(o_k) = 1$ if the event is true and zero otherwise, and $T(X_k)$ is the probability assigned to that outcome being true by some algorithm or forecaster. This linear transformation of the Brier score retains the same functional form as the original Brier score proposed by Brier (1950), and is strictly proper (Murphy & Winkler, 1970). Further, it has a straightforward interpretation where scores range from 0 to 100, with 100 being a perfect forecast over all events and maximally uninformed forecasts of $p = .5$ receive a score of 75. We use the transformed Brier score to assess the performance of each aggregation approach in this chapter.

In each experiment, we tested the crowd’s performance on forecasting problems from multiple domains and examined the effectiveness of weighting forecasters by their contribution from unrelated domains, relative to weighting by their contribution in the test domain. This comparison is captured by comparing the performance of two models: the ‘standard’ CWM, which derives weights using forecasters’ contribution using questions from the same domain as the test questions, and the cross-domain CWM (xCWM), which derives forecasters’ weights from their contributions on questions that are not from the same domain as the test questions. For each experiment, we investigated whether there was an overall difference in score between the xCWM and the CWM, and whether this difference was consistent across questions of each domain.

Additionally, we were also interested in whether cross-domain weighting would provide any benefits over single-question aggregation approaches, which do not require the elicitation of forecasters’ responses on questions with known outcomes. Thus, we compared the performance of

the xCWM to two single-question aggregation algorithms: (1) the Meta-Probability Weighting (MPW) algorithm, which has been found to be effective in the single-question domain (Martinie et al., 2020); and (2) a recalibrated version of the MPW algorithm (MPW-R), which has been shown to improve upon the forecasts of the MPW algorithm. The MPW-R algorithm transforms the predictions of the MPW algorithm by making them more extreme (i.e., closer towards the edges of the probability scale) using:

$$t(p_k) = \frac{p_k^a}{p_k^a + (1 - p_k)^a} \quad (4.3)$$

where p_k is the original aggregated probability forecast for the k th event, $t(p)$ are the recalibrated probabilities, and a is the recalibration parameter, which determines the strength of the transformation. This function, which has been used by Baron et al. (2014); Shlomi and Wallsten (2010); Turner et al. (2014); and others before them, extremises probability forecasts when $a > 1$ and anti-extremises when $0 < a < 1$. Baron et al. (2014) found that this function worked well in crowds containing experts at approximately $a = 2.5$, and in crowds containing non-experts, approximately $a = 3.5$. Since optimisation of this parameter value over training questions is not feasible in the single-question domain, we adopted the more conservative parameter value, 2.5, and applied this transformation to the forecasts from the MPW algorithm to obtain the forecasts of the MPW-R algorithm.

Martinie et al. (2020) found that while recalibration using this parameter value generally improved the performance of the MPW algorithm, the MPW-R algorithm is not necessarily guaranteed to outperform the standard MPW algorithm in all problem domains. By including both the MPW algorithm and the MPW-R algorithm, we provide a more advantageous comparison for single-question aggregation approaches over the xCWM. This comparison is important because we are interested in the improvement provided by the xCWM over algorithms that do not take into account forecasters' past performance. For example, if the xCWM provided little to no improvement over single-question aggregation approaches, then there would be no reason to elicit forecasters' responses to seed questions in unrelated domains. While we could potentially test a whole range of other single-question aggregation approaches for an even more comprehensive comparison (e.g., see Martinie et al., 2020; Palley & Soll, 2019; Prelec et al., 2017; Satopää et al.,

2016), past findings suggest that it is unlikely that these algorithms will outperform the standard or MPW-R algorithm (see Martinie et al., 2020), so we confined our attention to just these two algorithms. In a later section of this chapter, we show that our general findings hold even if we replaced the MPW and MPW-R algorithm with other single-question algorithms, such as the Minimal Pivoting approach (Palley & Soll, 2019) or weighting by Decision Similarity (Kurvers et al., 2019).

4.4 Experiment 1: Comparing cross-domain weighting to within-domain weighting and single-question aggregation approaches

Experiment 1 examined forecasters’ performance on two domains: American National Football League (NFL) trivia and general knowledge science trivia. Since the two domains require fundamentally different sets of knowledge and expertise, forecasters are unlikely to be experts simultaneously in both domains. We predicted that cross-domain weighting (i.e., the xCWM) would be generally less effective than within-domain weighting (i.e., the CWM). However, we expect that weighting by forecasters’ expertise on a large set of unrelated questions to be more effective than by aggregating forecasts using only forecasters’ responses on each question itself. Thus, we predicted that cross-domain weighting (i.e., xCWM) would be generally more effective than the two single question approaches – the MPW algorithm and the recalibrated MPW algorithm.

Additionally, we compared the performance of the cognitive model proposed by Lee and Danileiko (2014) to the performance of the xCWM and the MPW algorithm. The cognitive modeling approach requires forecasters’ responses over a large set of problems but does not require the true outcomes of those problems to be known to the decision maker. Thus, the cognitive modeling approach is more restrictive than single-question approaches, which do not require responses to more than one problem. However, the cognitive modeling approach can be applied to a wider range of problems than within-domain and cross-domain weighting approaches, since it does not require the outcomes of these training problems to be known. We therefore expect

the cognitive modeling approach to outperform single-question aggregation approaches but not cross-domain or within-domain weighting approaches.

4.4.1 Methods

We collected responses to 50 questions from the NFL domain and 50 questions from the science trivia domain. NFL questions were adapted from trivia questions on the www.funtrivia.com website, and then converted into true or false statements. Science Trivia questions were taken from the Grades dataset from Martinie et al. (2020), which comprised moderate difficulty general science questions from Biology, Chemistry, Geography, and Physics.

We recruited 30 respondents from Amazon Mechanical Turk; only respondents inside the US were able to participate in the experiment. Participants were paid a flat fee of \$4.00 for completing the survey. The survey was conducted on the Qualtrics platform. Before beginning the experiment, participants were first required to answer three basic logic questions, which we used to identify and exclude any non-human agents from the survey. Participants were then asked to answer each question as honestly as they could and without cheating (e.g., by looking up any of the questions online). Nine individuals who reported cheating at the task were excluded from the analyses; analyses were conducted on the data of the remaining 21 participants.

Participants completed 100 trials each, with each trial comprising one statement which was either true or false. Half the statements in each domain were true, and the other half were false. Participants were asked to provide their predictions about (1) whether the statement was more likely to be true or false, (2) the probability that they believed they were correct (from 50–100), and (3) what percentage of other forecasters would predict the statement to be true (from 0–100).

Participants were presented each question from the set of 100 questions in a randomised order. The list of questions used in this chapter are included in the Appendix (see section 6.3).

4.4.2 Analyses

We used the transformed Brier score discussed above as our measure of forecasting performance. Our main comparisons of interest were between the xCWM algorithm and the other algorithms, aggregated across all 100 questions in the dataset. Additionally, we calculated the mean scores of

each of the five algorithms within each of the two domains. As an additional descriptive statistic, in the cases where those algorithms also outperform the unweighted mean, we also report the improvement offered by xCWM over the unweighted mean compared to that offered by the MPW and MPW-R algorithms over the unweighted mean. This comparison captures the magnitude of improvement offered by xCWM over these single-question approaches.

We used the bias-corrected and accelerated bootstrap technique (Efron & Tibshirani, 1994) to compute 95% Confidence Intervals (CIs) with 10,000 samples for statistical inference.

Note that, since the goal of this chapter is to compare competing aggregation models, the validity of our statistical inferences therefore depends on the number of questions on which we compare these models, rather than the number of participants in the sample. Understandably, readers may be concerned that the relative performance of these aggregation approaches will not generalise to larger-sized crowds. However, in Experiment 4 below, we show that our results will generalise to crowd sizes of approximately 100 people.

We calculated these CIs both generally across all 100 questions in the dataset and separately for each domain. These CIs inform us about whether these differences are statistically significant at the $\alpha = 5\%$ level when the null hypothesis value is not contained in the interval (for all comparisons in this thesis $H_0 = 0$). By calculating the results separately for both the NFL and Science domains, we can identify whether this effect is robust across different environments.

One limitation of this statistical testing approach is that Frequentist Confidence Intervals are generally unable to provide evidence for the null hypothesis, since the results are conditioned on the fact that the null hypothesis is true (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Wagenmakers, Lee, Rouder, & Morey, 2019). Since we are interested in testing whether pairs of models are equivalent in performance (i.e., we are expecting evidence for the null hypothesis for some of these comparisons), another statistical testing approach – one that is able to identify evidence for the null hypothesis – is needed. For each of the statistical comparisons in this chapter, we therefore also report a Bayes factor calculated using a paired-samples Bayesian t-test in JASP (Wagenmakers et al., 2018). By convention, we used the default Cauchy prior with a scale parameter of .707. The Bayes factor provides an indication as to whether the null hypothesis or alternative hypothesis is better supported by the data. We interpret these Bayes factors in accordance with Kass and Raftery (1995), that is, any BF_{10} between 1 and 3 indicates weak

evidence favouring the alternative hypothesis; any BF_{10} between 3 and 20 indicates positive evidence; any BF_{10} between 20 and 150 indicates strong evidence; and any BF_{10} greater than 150 indicating very strong evidence. Bayes factors less than 0 were inverted, such that they indicated evidence for the null hypothesis (i.e., $BF_{10} = 1/BF_{01}$) using the same cutoff rules.

The predictions of each algorithm were obtained as follows:

1. For the unweighted mean algorithm, the prediction for each test question was calculated by averaging the probability forecast of all forecasters in the crowd for that question.
2. For the CWM algorithm (equation 4.1), we used leave-one-out cross-validation to estimate forecasters' weights from training questions of the same domain as the test questions. For example, for each of the 50 NFL questions, one was selected as a test question on a particular iteration of the cross-validation process and the remaining 49 NFL questions were used as training questions from which we computed forecasters' mean contributions. We then normalised forecasters' contributions (so that the total of all forecasters' contributions summed to one) and applied them as linear weights to forecasters' probability forecasts on the test question in order to generate an aggregated prediction. Forecasters whose contributions were negative were assigned weights of zero (see Budescu & Chen, 2015, , for details).
3. For the xCWM algorithm, weights were estimated on questions from the other domain. For example, for each of the 50 NFL questions used as a test question, the 50 Science Trivia questions were used as seed questions to derive forecasters' mean contributions, which we then applied as linear weights in the exact same way as we did for the CWM algorithm. Forecasters whose contributions were negative were assigned weights of zero. Since the test question was always from a separate domain from the training set, this meant that the exact same set of weights was used for all test questions in that domain.
4. For the MPW algorithm, predictions were generated using only the responses to the test question (i.e., it did not use any information or responses from other questions). The MPW algorithm linearly weights each forecaster's probability forecast on a test question by the absolute difference between the forecaster's probability forecast and their meta-prediction

about the average forecast of others, normalised by the sum of all absolute differences across all forecasters for that question (such that the total of all forecasters’ weights summed to one):

$$T_{MPW}(X_k) = \sum_{i=1}^{N_k} \frac{|P_{i,k} - M_{i,k}^P| P_{i,k}}{\sum_{j=1}^{N_k} |P_{j,k} - M_{j,k}^P|} \quad (4.4)$$

where $P_{i,k}$ is the probability forecast of the i th forecaster for the k th question, $M_{i,k}^P$ is that forecasters’ meta-prediction about the average probability forecasted of others. In experiments 1, 2, and 3, we elicited forecasters’ meta-predictions about the proportion of other forecasters voting “true” in each experiment. We thus make the assumption that forecasters’ responses to these two different types are similar enough that we can use one meta-prediction as a reasonable approximation of the other. In Experiment 4, we show that our results hold even when we elicit and use the proper meta-prediction responses from forecasters (i.e., meta-predictions about the average probability forecast made by other forecasters).

5. For the MPW-R algorithm, the MPW algorithm’s prediction was transformed using the recalibration function discussed above (equation 4.3 with a fixed parameter $a = 2.5$ (Baron et al., 2014), resulting in a forecast that is closer to the ends of the probability scale compared to the forecast of the MPW algorithm.
6. For the LD cognitive model, predictions were generated using a Bayesian hierarchical model, where parameters capturing forecasters’ expertise and calibration parameters were jointly estimated over the full set of responses. Forecasters’ observed responses reflect a perceived probability, offset by miscalibration, drawn from a Gaussian distribution centered on the latent true probability. The expertise parameter for each forecaster reflects the variance in the samples they draw and the calibration parameter for each forecaster reflects the extent to which they under-estimate large probabilities and over-estimate small probabilities. The expertise and calibration parameters for each forecasters are then aggregated to provide a model forecast for each event. Model predictions were generated using the code published online with the paper by the original authors. For an exact, technical specification of the

model, see Lee and Danileiko (2014).

4.4.3 Results

Figure 4.1 shows the mean performance of each algorithm separately for the NFL Trivia and Science Trivia domains. The xCWM and CWM achieved minimal differences in performance in both domains. While the xCWM performed significantly worse than the CWM in the NFL Trivia domain by 2.47 points (95% CI: [-3.77, -0.44], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 7.36$), there was no significant difference between their performance in the Science Trivia domain (95% CI: [-0.65, 1.73], with positive evidence in favour of the null hypothesis, $BF_{01} = 4.96$). Collapsing across all 100 questions, we found that overall there was no significant difference in performance between the xCWM and CWM (95% CI: [-1.95, 0.13], with weak evidence in favour of the null hypothesis, $BF_{01} = 1.693$).

The xCWM significantly outperformed both the MPW algorithm by 10.63 points (95% CI: [8.27, 13.04], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 8.813e^{10}$) and the MPW-R algorithm by 8.94 points (95% CI: [5.12, 13.78], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 230.3$). The difference in score between the xCWM and the unweighted mean (mean difference = 13.31) was about 5 times larger than the difference in score between the MPW algorithm and the unweighted mean (mean difference = 2.68) and about 3 times larger than the difference in score between the MPW-R algorithm and the unweighted mean (mean difference = 4.38). The xCWM thus provided better predictions than these single-question aggregation approaches by a substantial amount.

The LD cognitive model performed significantly worse than the xCWM algorithm in the NFL Trivia domain (95% CI: [-14.94, 9.40], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 3.61e^8$) and Science Trivia domain (95% CI: [-15.86, 8.20], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 44502.46$). On the other hand, the LD cognitive model did not significantly outperform the MPW algorithm in either the NFL Trivia domain (95% CI: [-4.70, 2.55], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.57$) or Science Trivia domain (95% CI: [-5.10, 1.05], with positive evidence in favour of the null hypothesis, $BF_{01} = 3.48$). Surprisingly, the LD cognitive model therefore performed slightly worse than the

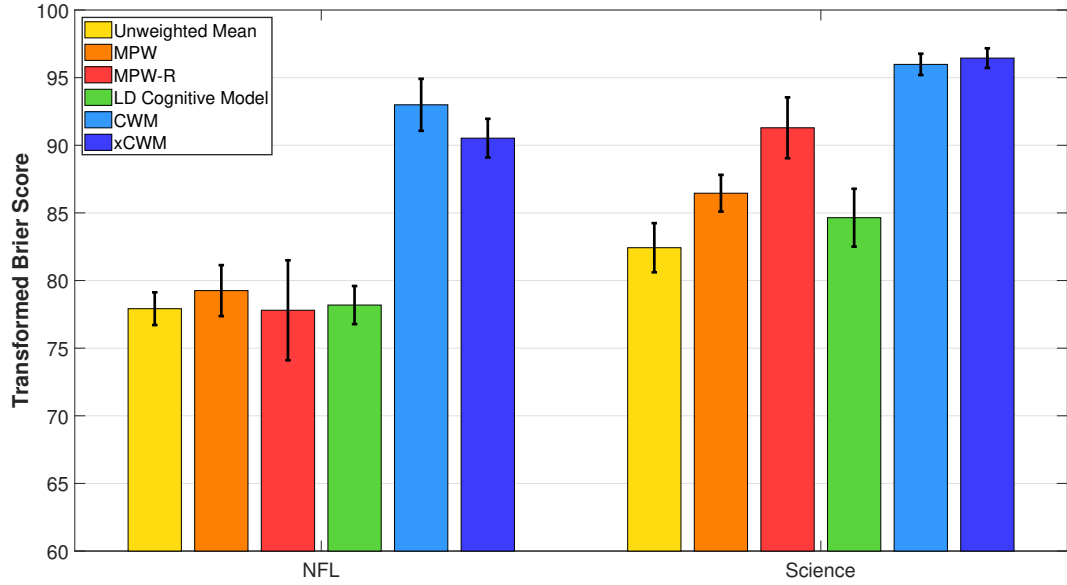


Figure 4.1: The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting algorithm, the Recalibrated Meta-Probability Weighting algorithm, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the NFL Trivia (left) and Science Trivia domain (right). Error bars show the standard error. The CWM significantly outperformed the xCWM in the NFL Trivia domain but not in the Science Trivia domain.

MPW algorithm, despite using a much larger set of forecasters' responses to produce a model forecast.

4.4.4 Discussion

We found that across 100 questions spanning the NFL Trivia and Science Trivia domains, there was no significant difference in performance between the xCWM and the CWM. Our results demonstrate that deriving contribution weights from seed questions of other domains was either just as effective as deriving contribution weights from questions from the same domain, or was only less effective by a very small margin. We did not expect cross-domain weighting to be so effective, since we had chosen questions relying on two distinctly different sets of knowledge. Despite this, we found that the domain from which we estimated forecasters' expertise made little difference to the performance of the final aggregated forecast.

Furthermore, we found that cross-domain weighting was significantly more effective than

single-question aggregation approaches and provided an increase in score over the unweighted mean at least three times larger than that provided by the single-question aggregation approaches. Similarly, the cross-domain weighting approach outperformed Lee & Danileiko’s (2014) cognitive modeling approach by a significant and substantial amount in both domains. The success of the cross-domain weighting approach has important practical implications about the choice of aggregation approach and decision makers’ choice to elicit forecasts about additional questions with known outcomes in order to estimate forecasters’ expertise.

In order to provide a more rigorous test of cross-domain weighting, the next experiment attempts to replicate these results using an additional domain of emotional intelligence, for which expertise should depend on an entirely different set of skills.

4.5 Experiment 2: Testing cross-domain weighting using performance on multiple unrelated domains

Experiment 2 tested an additional domain of emotional intelligence (EI), measured using questions adapted from the Situational Test of Emotional Understanding and Situational Test of Emotion Management (MacCann & Roberts, 2008) alongside the two domains of NFL Trivia and Science Trivia from Experiment 1. We chose to use these three domains because EI questions theoretically tap into a fundamentally different set of skills and characteristics compared to NFL Trivia and Science Trivia domains, which tap into domain-specific knowledge in general intelligence (e.g., Cattell, 1963; Lam & Kirby, 2002; Mayer, Salovey, Caruso, & Sitarenios, 2001). By choosing domains that are theoretically unrelated, we are able to provide a more rigorous test of the extent to which expertise can be accurately estimated from unrelated domains. We predict the xCWM weights derived from forecasters’ contributions on questions of EI to be much less effective when used to weight forecasts in the NFL and Science domains compared to weighting by contributions from the same domain. Additionally, we also predict that within-domain weighting on the EI questions (i.e., the CWM) would be much more effective than weighting by forecasters’ contributions on the NFL Trivia or Science Trivia domains or on both domains combined.

4.5.1 Methods

We repeated the same experiment as before but included an additional set of 50 trials where participants were presented with EI statements. The EI questions were adapted from the Situational Test of Emotional Understanding and Situational Test of Emotion Management developed and validated by MacCann and Roberts (2008). We chose to adapt questions from this source due to the fact the test relies on questions with objectively-correct answers rather than self-report scales with no objective answer. Additionally, test questions and answers were also provided online by the original authors (see MacCann & Roberts, 2008). 50 questions, originally in the form of four-alternative multiple choice, were randomly selected from these two tests and converted into two-alternative questions after removing two of the three possible incorrect options. While adapting these tests in such a way may reduce their validity as a measure of EI, these questions are still likely to be effective at capturing participants' expertise in the EI domain. Expertise on these adapted questions are unlikely to be due to information they have explicitly memorised beforehand, compared to the NFL Trivia and Science Trivia domains where this is likely to be the case.

In the same way as before, participants were asked to provide the votes, confidences, and meta-predictions to each statement. For the EI statements, participants were presented with two alternatives (as opposed to a choice between "True" and "False") and were then asked: (1) whether the first or second alternative was more likely to be correct, (2) the probability that they believed they were correct (from 50–100), and (3) their belief about the percentage of other forecasters who would choose the first alternative over the second alternative (from 0–100). The correct option for the EI questions were the first alternative on half of the trials (25 of the 50 trials), and the second alternative for the other half of the trials. Like before, these questions were presented in a randomised order and were randomly interspersed among the questions from other domains. Participants responded to 150 trials in total over 3 domains: 50 trials about NFL Trivia, 50 trials about Science Trivia, and 50 trials about EI. The list of questions used in this chapter are included in the Appendix (see section 6.3).

As before, we collected responses from 30 participants. Each participant was paid USD \$6.00 for completing the experiment. Only respondents inside the US were able to participate, and

participants from any of our previous experiments were prevented from participating in this experiment. Coincidentally, nine participants reported that they cheated during the experiment and were therefore removed from the analyses. The analyses were conducted on the remaining 21 people.

4.5.2 Analyses

Firstly, on the NFL and Science domains, we compared the performance of the CWM (i.e., within-domain weighting) to “EI Weights”, where forecasters’ test domain predictions were weighted by their contributions on the full set of 50 EI questions. We computed the 95% CI for mean difference in Transformed Brier Score between the CWM and EI Weights on each of the NFL Trivia and Science Trivia domains. As a test of the efficacy of cross-domain weighting relative to single-question approaches, we computed the difference in performance between (1) the EI Weights and the MPW algorithm and (2) the EI weights and the MPW-R algorithm. We computed the differences overall for the NFL and Science Trivia domains. All other algorithms were implemented in the same way as our earlier experiment.

Secondly, on the EI domain, we compared the performance of the CWM to “NFL Weights” and “Science Weights”, where we weighted forecasters’ EI predictions by their contributions in the NFL Trivia domain and Science Trivia domain, respectively. Additionally, we compared the CWM to the performance of the xCWM. The weights for the xCWM were calculated using forecasters’ contributions on both the NFL Trivia and Science Trivia domains combined. We note that by combining the other two domains results in twice as many questions in the training set compared to the training set for the CWM. However, we show in simulations in a later section of this chapter that our general findings hold even after ensuring both models use the same number of training questions. For the EI domain, we computed the 95% CI for mean difference in Transformed Brier Score: (1) between the CWM and NFL Weights, (2) between the CWM and Science Weights, (3) between the CWM and the xCWM, (4) between the xCWM and the MPW algorithm, (5) between the xCWM and MPW-R algorithm, (6) between the LD cognitive model and the xCWM, and (7) between the LD cognitive model and the MPW algorithm.

Using the same method as before, we also computed Bayes factors for each comparison.

4.5.3 Results

The NFL Trivia and Science Trivia domains

Figure 4.2 shows the mean performance of each algorithm separately on each of the NFL Trivia and Science Trivia domains. EI weighting performed worse than the CWM by a small margin in both the NFL Trivia domain (95% CI: [-2.52, 1.82], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.99$), and the Science Trivia domain (95% CI: [-1.85, 1.11], with positive evidence in favour of the null hypothesis, $BF_{01} = 6.00$). Thus, there was once again only a very small and non-significant difference in performance between within-domain weighting and cross-domain weighting.

EI weighting also significantly outperformed the MPW algorithm by 3.76 points (95% CI: [1.22, 6.30], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 5.16$) and the MPW-R algorithm by 5.67 points (95% CI: [1.49, 10.34], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 2.27$). Weighting forecasters' predictions in the NFL Trivia and Science Trivia domains by their past performance on a third, unrelated domain therefore outperformed single-question aggregation approaches by a substantial and significant amount.

The LD cognitive model performed significantly worse than EI weighting in the NFL Trivia domain (95% CI: [-19.82, -2.70], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 3.22$) and the Science Trivia domain (95% CI: [-24.40, -8.34], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 94.72$). Surprisingly, the LD cognitive model also performed significantly worse than the MPW algorithm in the NFL Trivia domain (95% CI: [-14.77, -1.70], with weak evidence in favour of the alternative hypothesis, $BF_{10} = 2.48$) and the Science Trivia domain (95% CI: [-19.08, -4.94], with positive strong in favour of the alternative hypothesis, $BF_{10} = 13.40$).

The Emotional Intelligence domain

Figure 4.3 shows the mean performance of each algorithm in the Emotional Intelligence (EI) domain. While once again, weighting by forecasters' contributions on other domains was slightly less effective than by contributions from the test domain, there was no significant difference in performance between the CWM and: NFL weights (95% CI: [-0.45, 5.24], with weak evidence in

favour of the null hypothesis, $BF_{01} = 1.62$), Science weights (95% CI: [-1.70, 2.14], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.58$), or xCWM weights (95% CI: [-1.75, 2.36], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.36$). Aggregated forecasts were therefore almost equally as accurate regardless of which domain’s contributions forecasters were weighted by.

The xCWM also significantly outperformed the MPW algorithm by 9.85 points (95% CI: [5.68, 14.26], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 64.97$) and the MPW-R algorithm by 15.09 points (95% CI: [7.62, 23.04], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 537.5$). Cross-domain weighting therefore outperformed single-question aggregation approaches by an even more substantial amount in the EI domain than in the other two domains.

The LD cognitive model performed significantly worse than the xCWM in the EI domain (95% CI: [-16.27, -7.35], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 3929.67$). There was no significant difference in performance between the LD cognitive model and the MPW algorithm (95% CI: [-5.86, 1.81], with positive evidence in favour of the null hypothesis, $BF_{01} = 4.06$).

4.5.4 Discussion

Despite using training questions from an entirely different domain we found that forecasters’ expertise on NFL and Science domains could be identified almost as effectively from contributions from EI questions as from their contributions on the NFL and Science domains. We found that forecasters’ contribution weights were highly effective regardless of the domain of the training questions that those weights were derived from. This was again unexpected, given that forecasters’ expertise on EI questions draw on a very different set of skills and characteristics compared to their expertise in NFL Trivia or Science Trivia. Furthermore, the cross-domain weighting approach seemed to outperform the cognitive modeling approach by a substantial margin, regardless of the domain. In conjunction with Experiment 1, our results here suggest that the cross-domain weighting approach can be highly effective across a range of forecasting problems.

Our current results also provide more nuanced insight into cross-domain weighting. As Figure

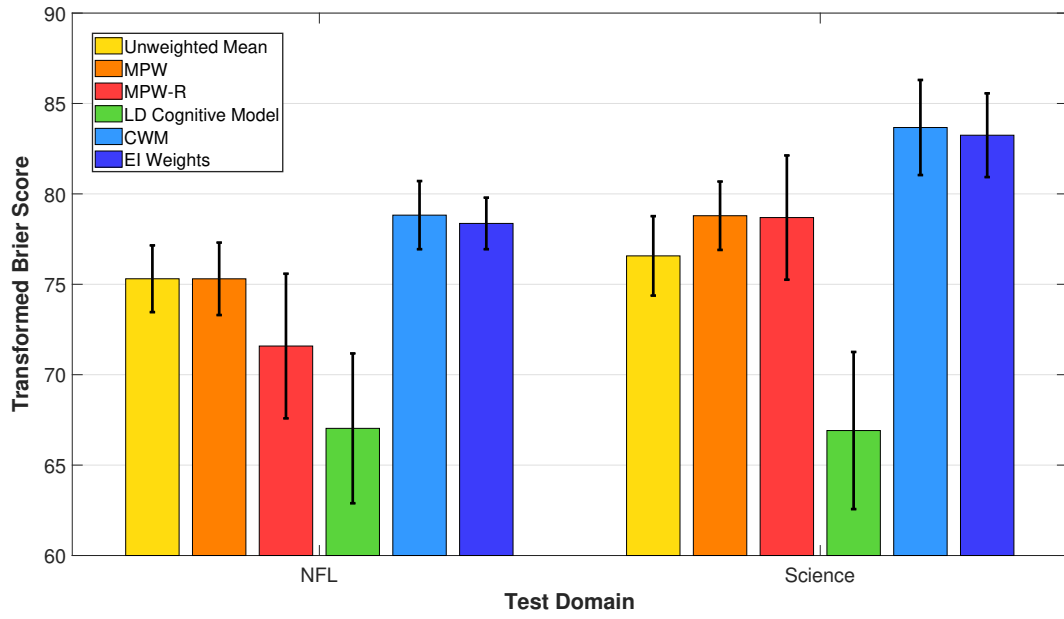


Figure 4.2: The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the NFL Trivia domain (left) and Science Trivia domain (right). Error bars show the standard error. There was no significant difference between the performance of the CWM and the Emotional Intelligence (EI) weights in either domain.

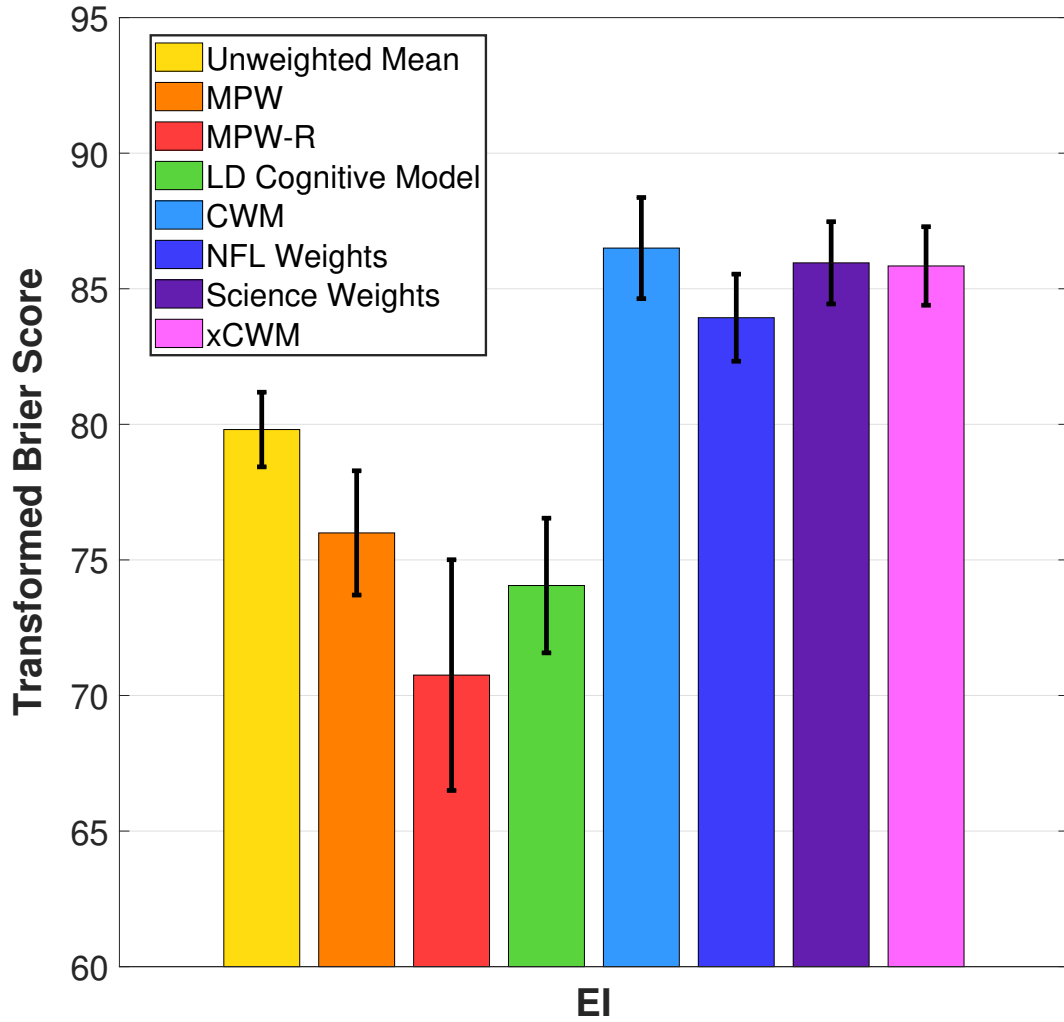


Figure 4.3: The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM), NFL Weights, Science Weights, and the cross-domain Contribution-Weighted Model (xCWM) on questions from the Emotional Intelligence (EI) domain. Error bars show the standard error. There was no significant difference between the performance of the CWM and NFL Weights, Science Weights, or the xCWM.

4.3 showed, cross-domain weighting may be slightly less effective on some domains than others (e.g., applying NFL weights in the EI domain). In practice, it is unlikely that performance in unrelated domains will always predict forecasters’ performance on a domain of interest. However, as shown by the xCWM in Figure 4.3, cross-domain weighting can be made more robust by combining forecasters’ performance on multiple unrelated domains. We therefore adopted this combined-domain approach for the xCWM for the remaining experiments.

To provide a stronger test of these results, we once again attempted to replicate our findings using more distinct and unrelated domains.

4.6 Experiment 3: Art, Science Trivia, and Emotional Intelligence problems

In Experiment 3, we sought to use even more distinct domains by replacing the questions of NFL Trivia with questions requiring judgment of prices of artworks. We used artwork prices because judgements about their price require a fundamentally different set of skills compared to answering questions from the Science Trivia or EI domains. We selected artworks whose prices did not depend directly on the complexity, size, or other perceptual dimensions of the artwork. Expertise in the judgements of these artwork prices therefore most likely required forecasters to have previous knowledge about the different styles of works by a wide range of professional artists, and to be able to recognise these artworks visually. In contrast, expertise in the Science Trivia and EI domains could be obtained simply by having paid more attention to science classes in high school or developed through forecasters’ social interactions with others. Thus, we would expect forecasters’ expertise on the Art domain to be much better predicted by their performance on other art judgements in the same domain compared to their performance on questions in the other domains. Similarly, we would expect forecasters’ expertise on the Science Trivia and EI domains to be better predicted by their performance within the same domain than by their performance on other domains.

4.6.1 Methods

We applied the same methodology from the previous experiment, but replaced the set of questions from the NFL Trivia domain with a set of questions where participants were asked to judge the value of different artworks. On each of the Art trials, participants were presented with an image of an artwork and asked to judge whether the market price of the original version of that artwork would exceed USD \$10,000. They were then asked to respond to provide their votes, confidence and meta-predictions in the same way as before.

The artworks used were taken from the online websites listing professional artworks along with their prices (e.g., Sotheby auctions), and online websites that list famous historical artworks. Images of artworks worth less than USD \$10,000 were obtained from online websites selling original amateur artworks (e.g., Etsy), and after these ‘cheaper’ artwork images were selected, they were double-checked using Google reverse-image search to ensure that they were not sold or listed on any website for more than USD \$10,000. The list of questions used in this chapter are included in the Appendix (see section 6.3).

We reduced the number of questions from each domain to 40 per domain. Participants therefore completed 120 trials in total. All 120 questions were presented in a randomised order. Each participant was paid USD \$4.50 for completing the experiment. Participants from any of our previous experiments were prevented from participating in this experiment. We collected responses for 30 participants, and we then excluded participants who reported cheating during the experiment as we have for the previous experiments. One participant who failed to complete the survey and eight other participants who reported cheating were excluded from the analyses. Analyses were conducted on the remaining 21 participants.

4.6.2 Analyses

We repeated the same set of analyses, with all models and algorithms implemented in the same way as before.

4.6.3 Results

Figure 4.4 shows the mean performance of each algorithm separately on each of the three domains. Surprisingly, the xCWM significantly outperformed the CWM overall by 1.76 points (95% CI: [0.91, 2.71], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 87.51$). The xCWM also significantly outperformed the CWM for two of the three individual domains: Science Trivia by 1.36 points (95% CI: [0.28, 2.52], with weak evidence in favour of the null hypothesis, $BF_{01} = 1.97$), EI by 3.63 (95% CI: [1.77, 5.77], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 32.6$), but not Art (95% CI: [-1.03, 1.53], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.33$).

Not surprisingly, the xCWM also significantly outperformed both the MPW algorithm by 5.84 points (95% CI: [2.78, 8.91], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 63.25$) and the MPW-R algorithm by 12.79 points (95% CI: [7.54, 18.19], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 2890.93$). The xCWM therefore provided better predictions than any single-question algorithm by an amount much larger than what we have seen in the previous experiments.

The LD cognitive model performed significantly worse than the xCWM in all three domains: Art (95% CI: [-22.65, -3.52], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 3.02$), Science Trivia (95% CI: [-26.71, -8.55], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 60.07$), and EI (95% CI: [-12.55, -0.51], with weak evidence in favour of the alternative hypothesis, $BF_{10} = 1.14$). The LD cognitive model also performed significantly worse than the MPW algorithm in the Science Trivia domain (95% CI: [-14.73, 4.97], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 16.44$), but was not significantly better or worse in the Science Trivia domain (95% CI: [-23.14, -5.80], with positive evidence in favour of the null hypothesis, $BF_{01} = 3.89$), or EI domain (95% CI: [-8.38, 7.74], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.86$).

4.6.4 Discussion

Our results showed even stronger evidence for the efficacy of the cross-domain weighting approach than the results from the previous two experiments. We found that cross-domain weighting

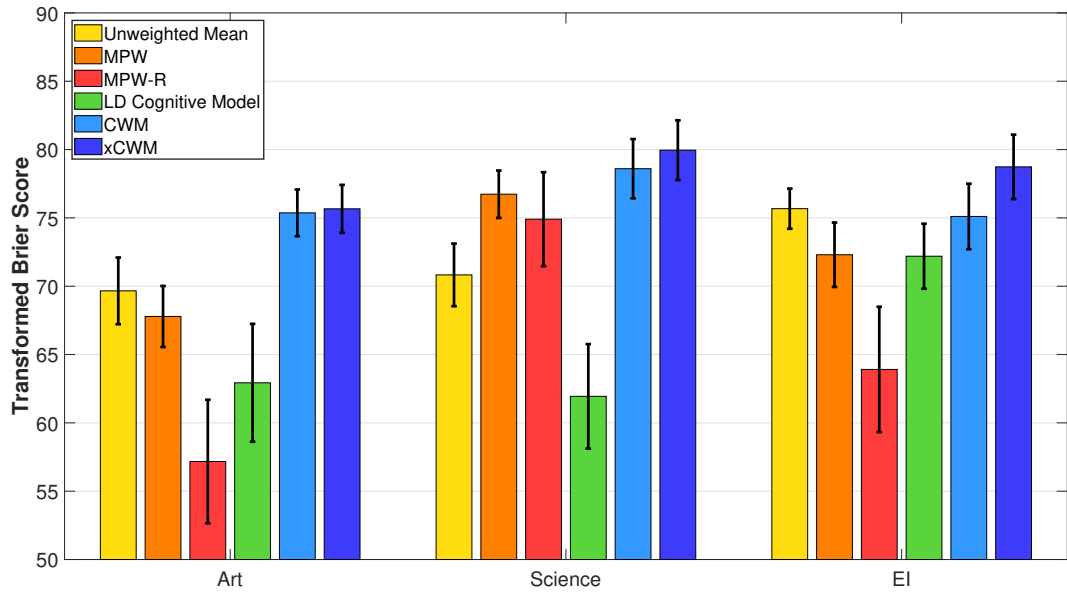


Figure 4.4: The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence questions (right). Error bars show the standard error. The xCWM significantly outperformed the CWM in the Science Trivia and Emotional Intelligence domains, but not in the Art domain.

could outperform weighting by past performance in the same domain – a result which is entirely novel. While the improvement offered by cross-domain weighting was relatively small, it was statistically significant both generally, and specifically on two out of the three domains. Combined with our previous results, these results provide strong evidence that forecasters’ expertise can be accurately estimated for a wide range of forecasting problems by using forecasters’ performance on questions from unrelated domains. Furthermore, cross-domain weighting was substantially more effective than both single-question aggregation approaches as well as the cognitive modeling approach proposed by Lee and Danileiko (2014). Our results therefore highlight the benefit of estimating cross-domain weights over simply estimating forecasters’ expertise using single-question approaches and cognitive modeling approaches.

One potential explanation for these results is that the contribution metric may be unreliable under modest sample sizes (Chen et al., 2016), and these single-question aggregation approaches may be even less robust. Our Experiments 1, 2, and 3 used modest sample sizes with fewer than 30 people. The poor performance of these single-question approaches could potentially be attributed to the relatively small sample size. We therefore sought to replicate our results using the same set of questions but with a much larger sample size.

4.7 Experiment 4: Replicating Experiment 3’s results using a larger sample

Experiment 4 replicates the results from Experiment 3 using a much larger sample size. Chen et al. (2016) showed that the contribution metric is less robust under small sample sizes, so larger sample sizes will allow us to draw stronger conclusions. Furthermore, the use of a small sample size may be particularly disadvantageous for single-question weighting approaches, which rely on stability in forecasters’ responses (i.e., their forecasts and meta-predictions) on individual questions. This is less reliable than forecasters’ past performance on a set of other questions, which are validated against an objective outcome. In contrast to previous studies on the MPW and MPW-R algorithm (Martinie et al., 2020), which used sample sizes of approximately 100 people, our Experiments 1, 2, and 3 only used samples of approximately 30 people. By replicating

our previous results using a larger sample size, we would therefore be able to provide a more rigorous comparison of the cross-domain weighting approach to both standard within-domain weighting and single-question aggregation approaches. Our results here would also be able to demonstrate that the efficacy of cross-domain weighting is not limited to small sample sizes.

The current experiment also addresses an important limitation from Experiments 1, 2, and 3. Earlier in the chapter, we noted that in order to apply the MPW and MPW-R algorithms, we assumed that forecasters' meta-predictions about the proportion of votes for each outcome by other people could be treated as equivalent to their meta-prediction about the average probability forecasted by others. Martinie et al. (2020) had previously shown that this may disadvantage these single-question aggregation approaches since they theoretically operate on the latter type of meta-prediction. In the current experiment, we elicited forecasters' meta-predictions about the average probability forecast of others rather than their meta-predictions about the proportion of votes of others. This change provides a fairer comparison for the MPW and MPW-R algorithms, and therefore eliminates any alternative explanations for the extent to which the xCWM outperforms these single-question approaches.

4.7.1 Methods

We applied the same methodology from the previous experiment with the main difference being a larger sample. In the previous three experiments, we elicited forecasters' votes, confidences, and a meta-prediction about the proportion of votes other people voting for the first option. In the current experiment, we instead presented forecasters with two options and then elicited forecasters' estimates of (1) the probability that option A is correct, and (2) the average probability forecasted by others. The set of questions in this experiment was identical to that of Experiment 3. The list of questions used in this chapter are included in the Appendix (see section 6.3).

Participants completed 120 trials in total. Each participant was paid USD \$4.00 for completing the experiment. Participants from any of our previous experiments were prevented from participating in this experiment. We collected responses for 100 participants, and we then excluded participants who reported cheating during the experiment as we have for the previous experiments. 12 participants were excluded and analyses were conducted on the remaining 88 people.

4.7.2 Analyses

We repeated the same set of analyses as before. Our main comparison is between the mean transformed Brier score of the xCWM algorithm and the CWM algorithm, aggregated across all 120 questions in the dataset. As before, we computed 95% CIs for the mean difference in score between the xCWM and the CWM, both generally across all 120 questions and specifically for the 40 questions in each of the three domains. We tested whether the xCWM generally outperformed the MPW algorithm and the MPW-R algorithm, and examined the difference in improvement offered by the xCWM relative to these single-question aggregation approaches. We also compared the performance of the cognitive model proposed by Lee and Danileiko (2014) to the xCWM and the MPW algorithm. All algorithms were implemented in the same way as before. As before, we also computed a Bayes factor for each statistical comparison.

4.7.3 Results

Figure 4.5 shows the mean performance of each algorithm separately on each of the three domains. There was no significant difference in performance between the xCWM and CWM overall across the 120 questions (95% CI: [-1.40, 0.08], with weak evidence in favour of the null hypothesis, $BF_{01} = 2.52$), or for any of the three individual domains: Art (95% CI: [-0.88, 1.64], with positive evidence in favour of the null hypothesis, $BF_{01} = 5.20$), Science Trivia (95% CI: [-2.36, 0.08], with weak evidence in favour of the null hypothesis, $BF_{01} = 1.63$), or EI domains (95% CI: [-2.62, 1.42], with weak evidence in favour of the null hypothesis, $BF_{01} = 1.49$). Once again, there was a small and non-significant difference in score between the xCWM and CWM.

The xCWM significantly outperformed both the MPW algorithm by 2.55 points (95% CI: [1.00, 4.05], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 12.27$) and the MPW-R algorithm by 3.76 points (95% CI: [0.86, 7.00], with weak evidence in favour of the alternative hypothesis, $BF_{10} = 1.63$). Furthermore, the improvement offered by the xCWM over the unweighted mean was substantially larger than the improvement offered by the MPW algorithm. Specifically, the difference in score between the xCWM and the unweighted mean (mean difference = 4.29) was about 2.5 times larger than the difference in score between the MPW algorithm and the unweighted mean (mean difference = 1.75). This difference was even more

favourable for the xCWM when compared to the MPW-R algorithm, which performed worse than the standard MPW algorithm. The xCWM therefore once again provided better predictions than any single-question algorithm by a substantial amount.

The LD cognitive modeling approach performed significantly worse than the xCWM in all three domains: Art (95% CI: [-35.84, -6.11], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 4.48$), Science Trivia (95% CI: [-48.23, -22.03], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 4382.07$), and EI (95% CI: [-39.74, -13.09], with strong evidence in favour of the alternative hypothesis, $BF_{10} = 65.46$). The LD cognitive modeling approach also performed significantly worse than the MPW algorithm in all three domains: Art (95% CI: [-34.35, -2.04], with weak evidence in favour of the alternative hypothesis, $BF_{10} = 1.32$), Science Trivia (95% CI: [-48.29, -20.02], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 1605.27$), and EI (95% CI: [-39.49, -8.09], with positive evidence in favour of the alternative hypothesis, $BF_{10} = 5.78$). This result is surprising, given that the LD cognitive model generates forecasts using forecasters responses over a large set of questions – not just a single question.³

4.7.4 Discussion

Results from this experiment provide further evidence of the efficacy of cross-domain weighting. We found that forecasters’ expertise, measured as their contributions, could be estimated almost as effectively from unrelated domains as from questions in the same domain. We did not find any evidence that weighting by cross-domain expertise, captured by the performance of the xCWM, was significantly less effective than weighting by within-domain expertise, as captured by the performance of the CWM, for any of the three domains. Given that expertise in the judgement of art prices, science trivia, and EI questions rely on fundamentally different sets of skills and knowledge, our results suggest that forecasters’ contributions can serve as a generalised metric of expertise. Furthermore, these results suggest that the relationship between the test and training domains play a much smaller role than we would typically believe in our ability to

³We initially suspected that the model was not fitting correctly due to mis-specified parameter values for the prior distributions. However, inspection of the Markov Chain Monte Carlo (MCMC) chains revealed that the model was converging quickly and precisely to the same parameter values, regardless of the parameter values specified for the prior distributions. This suggests that the model was fitting correctly and the MCMC samples were good approximations to the posterior distribution for each parameter.

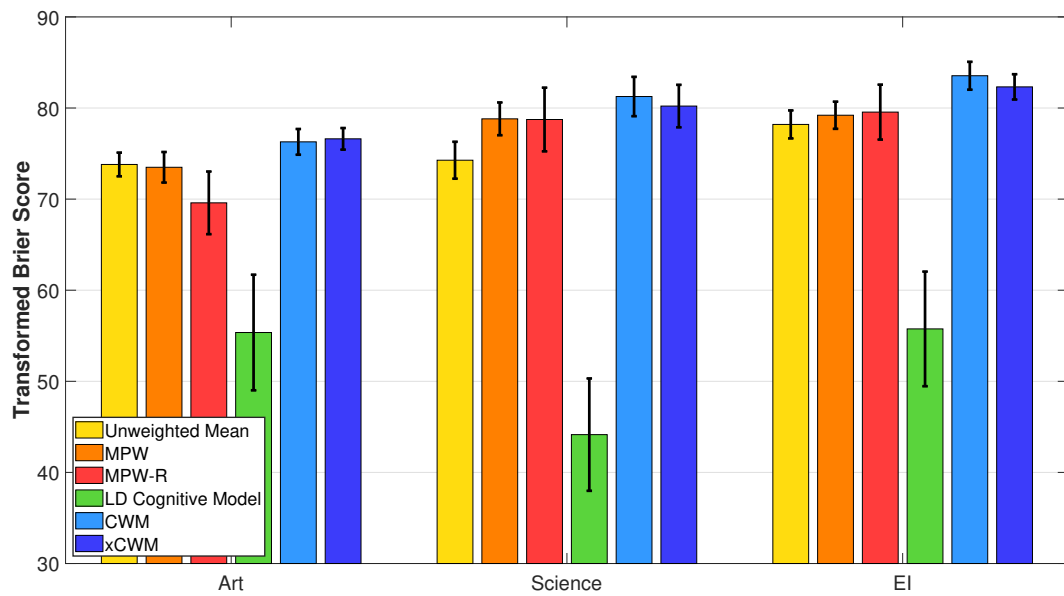


Figure 4.5: The mean transformed Brier score for the Unweighted Mean, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the LD cognitive model, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence (EI) questions (right). Error bars show the standard error. There was no significant difference between the performance of the xCWM and CWM in any of the three domains.

estimate forecasters’ expertise.

Our results also indicate that the cross-domain weighting approach can provide substantial improvements in forecasting performance over single-question aggregation approaches. We were able to quantify the improvement offered by the xCWM over single-question aggregation approaches, such as the MPW algorithm. This result therefore captures the trade-off in performance between eliciting responses from forecasters on an unrelated set of questions with known outcomes and simply estimating their expertise using single-question approaches.

Lastly, we have demonstrated the improvement in performance offered by both single-question aggregation approaches and cross-domain weighting approaches offered by a cognitive modeling approach developed by Lee and Danileiko (2014). Although the cognitive modeling approach requires forecasters’ responses over a large set of questions, it does not seem to provide any advantage over single-question aggregation approaches, which require less information from forecasters and also produce more accurate probabilistic forecasts.

4.7.5 Post-hoc Simulations

Another potential explanation for why cross-domain weighting was just as effective as within-domain weighting could be that the xCWM was trained on a larger set of questions, therefore enabling a more accurate estimate of forecasters’ contributions from these other domains than from the same domain. The training set size for the xCWM algorithm in Experiment 2, 3, and 4 was over twice as large as the training set size for the CWM algorithm. For example, in Experiment 4, the xCWM weights for the 40 questions in the Art domain was calculated using forecasters’ average contribution on the 80 questions from the Science Trivia and EI domains. In contrast, the CWM weights for those same 40 Art questions were estimated using a leave-one-out cross-validation (i.e., the jack-knife) procedure, and therefore used 39 training questions. It is possible that the performance of the xCWM is partly attributable to the difference in training set size used by the two algorithms. To test whether this explanation can account for the impressive performance of cross-domain weighting, we conducted additional tests using post-hoc simulations where we varied set size of training data and compared the performance of the xCWM to that of the CWM.

We simulated the change in mean transformed Brier score for the CWM and xCWM over different sized training sets by using the empirical bootstrap (Efron & Tibshirani, 1994) to resample data from each of the four experiments. For each test event, we divided forecasters’ probability forecasts on training questions into (1) a pool for questions from same domain as the test question, or (2) a pool for questions from other domain(s). To ensure that both within-domain and cross-domain pools contained the same number of training questions before applying the bootstrap, on each bootstrap iteration a subset of questions equal to the number of training questions in the within-domain pool was randomly selected without replacement from the cross-domain pool of training questions. Forecasters’ responses to the remaining questions in the cross-domain pool were removed for that iteration and this ensured both CWM and xCWM training sets were drawn from pools of the exact same size. A random set of k training questions was then randomly sampled with replacement from each pool. We repeated this process 1000 times for each training set size in the range of [10, 20, 30, 40, 50] for data from Experiments 1 and 2, and [10, 20, 30, 40] for Experiments 3 and 4. On each of the 1000 iterations, we calculated the performance of the CWM and xCWM, averaged across the 1000 iterations to obtain a single score for each algorithm on that test event, calculated using k training questions. We then averaged the performance of both algorithms across each test event (i.e., and across every domain) in the dataset to obtain their mean score for that dataset, and repeated this for each of the four experiments. As a reference, we also calculated the mean score of the unweighted mean on the original sample in each dataset. As an inferential test regarding the difference in score between the CWM and xCWM, we compared the general performance of the two algorithms over all four datasets using the largest training set size for the simulations from each respective dataset.

Figure 4.6 shows the mean score of the CWM and the xCWM across different training set sizes for each dataset. We can see that in all four figures, the difference in score between the CWM and xCWM was fairly consistent regardless of training set size. The difference in score between the CWM and xCWM appeared to be particularly small on smaller set sizes. These results therefore suggest that our results demonstrating the efficacy of cross-domain weighting can be expected to generalise to cases with fewer training events.

More importantly, we can see there was very little difference in score between the CWM and xCWM in all four datasets, regardless of the number of training events used. While our original

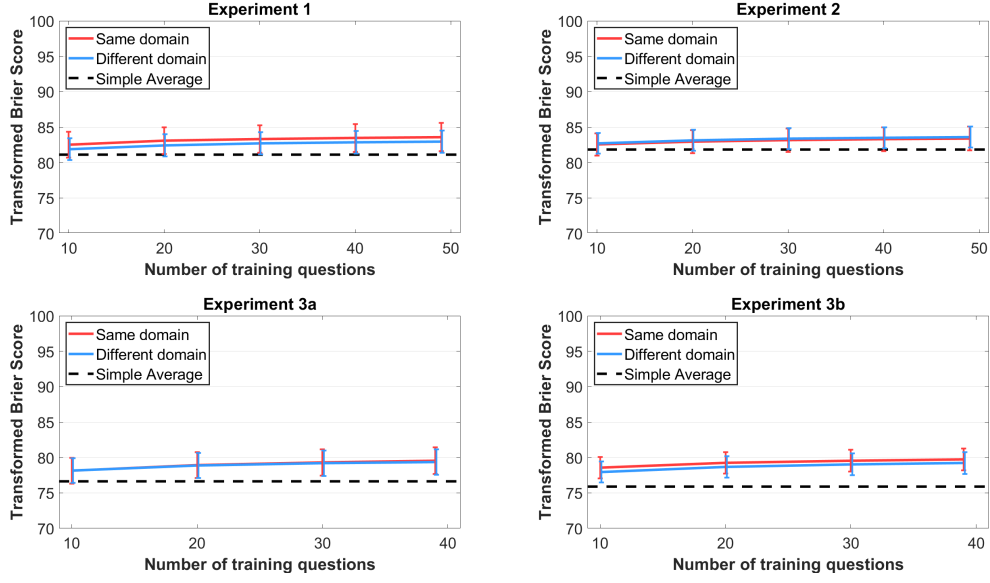


Figure 4.6: Simulation using results from Experiment 1-4 showing the mean transformed Brier score for the Contribution-Weighted Model (red), which is trained on questions from the same domain as the test domain, compared to the cross-domain Contribution-Weighted Model (xCWM), which is trained on questions from a different domain to the test domain. Performance is calculated across all questions in each experiment, shown as a function of training set size. The performance of the simple average (dashed line), which does not use training data, is shown for reference.

results may have benefited the xCWM by using more training events than the CWM to estimate forecasters' contributions, our results here suggest that the benefit this would have provided would be very small. Unsurprisingly, after recomputing the 95% CIs for mean difference in score across all 490 questions in the four experiments using the largest training set size for each experiment, we found no statistically significant difference between the score of the CWM and xCWM (95% CI: [-0.34, 0.49], with positive evidence in favour of the null hypothesis, $BF_{01} = 18.31$).

Overall, these simulation results suggest that the efficacy of cross-domain weighting could not be attributed to difference in training set size in any of our previous results. Furthermore, our results suggest that the forecasters' contributions estimated from questions from other domains are robust under different training set sizes. These results are highly consistent with those of Chen et al. (2016), who demonstrated the robustness of the standard CWM across different training set sizes. Our results are novel in showing that this robustness extends across forecasters' contributions estimated using questions from unrelated domains.

4.7.6 Comparing cross-domain weighting to other single-question aggregation approaches

In this section, we compare the performance of the xCWM to two other single-question aggregation approaches: the Minimal Pivoting approach (Palley & Soll, 2019) and weighting by forecasters' average Decision Similarity (Kurvers et al., 2019). We applied these models using the data collected in Experiment 4, alongside the other algorithms we tested in Experiment 4. We computed 95% CIs for the overall difference in Transformed Brier Score between the xCWM and (1) Decision Similarity Weighting model and (2) Minimal Pivoting model across the three domains.

Figure 4.7 shows the performance of algorithm in each of the Art, Science Trivia, and Emotional Intelligence domains. The xCWM significantly outperformed Decision Similarity Weighting by 4.64 points (95% CI: [2.52, 6.63], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 597.39$) and Minimal Pivoting by 2.61 points (95% CI: [1.39, 3.81], with very strong evidence in favour of the alternative hypothesis, $BF_{10} = 431.86$). The MPW algorithm offered the best performance out of all four single-question aggregation approaches, although there was no significant difference in performance between the MPW algorithm and the Minimal Pivoting model (95% CI: [-1.60, 1.71], with positive evidence in favour of the null hypothesis, $BF_{01} = 9.83$). Our results here justify our choice of the MPW algorithm and MPW-R algorithm to represent single-question approaches in the main text. These results also reinforce our finding that cross-domain weighting, on average, outperforms single-question aggregation approaches by a substantial margin.

4.8 General Discussion

The aim of the current chapter was to investigate whether forecasters' expertise could be accurately identified using their performance on decision problems in unrelated domains. We quantified forecasters' expertise by their contributions to the crowd forecast using the contribution-weighted model (CWM; Budescu & Chen, 2015), which we then used to weight forecasters' predictions. We estimated forecasters' contributions from either forecasters' responses to questions within the same domain or from their responses to questions from other domains, comparing two models:

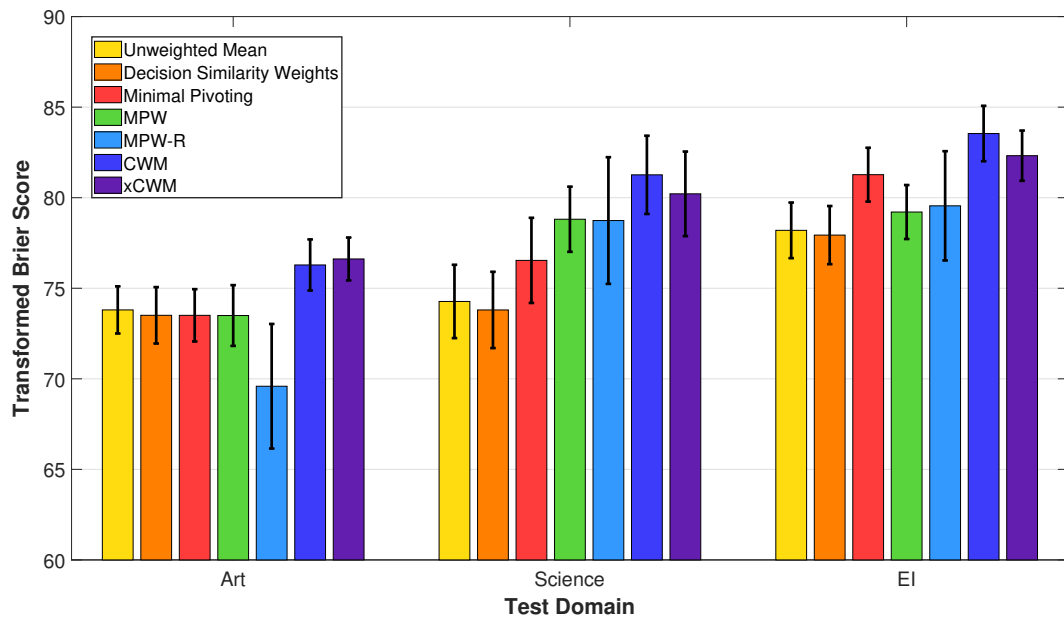


Figure 4.7: The mean transformed Brier score for the Unweighted Mean, Decision Similarity Weighting, the Minimal Pivoting approach, the Meta-Probability Weighting (MPW) algorithm, the Recalibrated Meta-Probability Weighting (MPW-R) algorithm, the Contribution-Weighted Model (CWM) and the cross-domain Contribution-Weighted Model (xCWM) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence (EI) questions (right). Error bars show the standard error.

the CWM, where forecasters’ contributions were estimated using their responses to questions within the same domain, and the cross-domain CWM (xCWM), where forecasters’ contributions were estimated using their responses to questions from other domains. Over four experiments, we examined the performance of the xCWM relative to the CWM using several theoretically distinct question domains. Additionally, we investigated the performance of cross-domain weighting compared to two single-question forecast-aggregation approaches: the Meta-Probability Weighting (MPW) algorithm and the Recalibrated MPW algorithm (Martinie et al., 2020). We also compared the performance of cross-domain weighting and the MPW algorithm to that of the cognitive modeling approach proposed by Lee and Danileiko (2014).

Our results across four experiments showed very little difference in forecasting performance between the xCWM and CWM. Cross-domain weighting further outperformed single-question aggregation approaches in all four experiments. Our results were consistent across small and large crowds of forecasters and across different problem domains, including sporting NFL trivia, general science trivia, judgements of fine art, and questions testing emotional intelligence. The generality of our findings is demonstrated by the consistency in our results across a wide range of problem domains.

Interestingly, across each experiment, the cognitive modeling approach proposed by Lee and Danileiko (2014) performed significantly worse than the xCWM. The LD cognitive modeling approach also performed consistently worse than the MPW algorithm for the majority of domains, although for many of these comparisons the difference in performance was not significant. This result is surprising, given that the LD cognitive model requires forecasters to provide forecasts to a large set of questions, whereas these single-question aggregation approaches only require forecasters’ responses to a single question. These results suggest that the assumptions of the LD cognitive model may not be appropriate in many of these domains, particularly in those where the majority of forecasters may be biased. In these cases, forecasters’ probabilistic forecasts may not be centred correctly on the true probability for each event, and the model may generate poor estimates for forecasters’ expertise and calibration under these conditions. It would be interesting for future research to test the efficacy of other cognitive modeling approaches against the single-question and cross-domain weighting approaches, to determine the extent to which these results generalise.

Our results have demonstrated the efficacy of cross-domain weighting relative to within-domain weighting and single-question aggregation approaches. Previous applications of other expertise-identification approaches such as the CWM have been largely limited to estimating forecasters’ expertise by their performance within similar or identical domains (Budescu & Chen, 2015; Cooke, 1991; Mellers et al., 2015). While the CWM’s robustness across different crowd sizes, crowd compositions, and number of training questions have been demonstrated previously (Budescu & Chen, 2015; Chen et al., 2016), no study to date has examined the extent to which forecasters’ contributions can be estimated from their performance on unrelated domains. Here, we have shown that forecasters’ contributions can be estimated effectively using their performance on unrelated domains, compared to their performance on similar domains. We have therefore demonstrated the contribution-metric to be even more versatile for extracting and identifying expertise than what has been shown in the existing literature.

Our results also show that cross-domain weighting is consistently and substantially more effective than single-question aggregation approaches. While previous research has shown that single-question aggregation approaches can be useful for identifying and extracting expertise when forecasters’ performance on the relevant domains are unknown (Martinie et al., 2020; Prelec et al., 2017), the current results suggest that expertise can be effectively estimated from unrelated domains – much more effectively than from single-question approaches. As our simulations show, forecasters’ expertise can be estimated effectively with as few as 20 training questions, regardless of domain, providing an improvement over simple averaging that is consistently several times better than that provided by single-question aggregation approaches. The cross-domain weighting approach is therefore an attractive and effective alternative for decision makers seeking to improve forecasts on novel problems for which there are no records of forecasters’ expertise.

The results observed here have interesting implications for the psychological processes determining performance on different problem domains. As Mellers et al. (2015) discussed, forecasters’ performance on forecasting problems are often driven by a range of factors: forecasters’ cognitive abilities and styles, task-specific skills, motivation and commitment, and access to enriched environments. The finding that forecasters’ expertise can be estimated effective across domains suggests that their performance is largely due to general factors such as cognitive ability and styles, motivation and commitment, and potentially access to enriched environments, rather than

task-specific skills. Given the wide range of problems examined in this chapter, this appears to be quite a general result. Nonetheless, these results may be limited by the fact that we selected decision problems that could be completed easily by laypeople. It's unlikely that these results will generalise to problems where task-specific expertise are expected to play a much larger role in determining forecasting performance.

4.8.1 Conclusions

Our primary focus in this chapter was not to find ways to maximise forecasting performance, but to demonstrate the efficacy of cross-domain weighting. Future studies may wish to investigate extensions to cross-domain weighting to maximise forecasting performance, for example, by combining cross-domain weighting with recalibration (e.g., Baron et al., 2014; Turner et al., 2014). It remains unclear, however, whether recalibration approaches would be as effective across domains, since unrelated domains can differ drastically in their difficulty, and question difficulty been previously shown to be important for determining recalibration efficacy (Martinie et al., 2020). Excluding more forecasters may also be an effective way at improving the xCWM forecast, as demonstrated for the CWM (Chen et al., 2016). By excluding forecasters who have positive but small contributions to the crowd, it may be possible to obtain a crowd of fewer but better-performing experts, thus improving the xCWM forecast. While our simulations suggest that the xCWM and CWM have similar levels of robustness to a reduction in training set size, it remains unclear the xCWM is as robust as the CWM to these other types of changes. Thus, we hope our results will inspire future researchers to examine the efficacy of cross-domain weighting approaches more generally, for example, by applying other advanced forecast aggregation approaches, or by testing our findings in other types of forecasting problems.

5 | Concluding Remarks

This thesis examined a number of approaches for improving aggregated predictions in the single-question and multi-question domains. The primary contribution of this thesis was to develop a method for identifying and leveraging the latent expertise of individual forecasters in the crowd in binary decision problems where forecasters' expertise cannot be easily identified. Existing aggregation methods that do not require information about forecasters' past performance do not adequately account for the difference in the expertise between forecasters. Existing approaches could therefore be improved by better accounting for these differences. This thesis provided a comprehensive demonstration of how forecasters' meta-predictions about the predictions made by other forecasters can be used to generate accurate forecasts in both categorical and probabilistic forecasting problems across a range of domains: predicting the correct answers to general knowledge, science, and sporting trivia; the classification of skin lesions; judgements about the prices of artworks; forecasts about the outcomes of football matches; and questions from a test of emotional intelligence. In the following paragraphs, we summarise the main contributions of the key chapters in this thesis and discuss the implications and future considerations for this research.

In Chapter 2, we proposed a novel reformulation of the Surprisingly Popular (SP) algorithm, in which forecasters' votes are weighted by the absolute difference between their votes and meta-predictions about the average vote of others. We examined differences in the weights assigned by the SP algorithm to different subsets of forecasters in the crowd and showed that the SP algorithm generally assigns greater weight to high-performing individuals than low-performing individuals. Based on this and the theoretical model developed in Wilkening et al. (2020), we proposed that a mechanism of the SP algorithm is to leverage the latent expertise in the crowd. We tested this

mechanism using a large dataset that varied in the level of latent expertise in the crowd over five levels, and found that, consistent with the mechanism we proposed, the SP algorithm offered the greatest advantage over other algorithms when crowds comprised a heterogeneous mixture of both experts and novices.

In Chapter 3, we tested the efficacy of different single-question aggregation approaches on probabilistic forecasting problems and showed how a meta-probability weighting approach, where forecasters' probabilistic forecasts are weighted by the absolute difference between their probability forecast and their meta-prediction about the average probability forecast of others, provides a powerful measure of forecasters' expertise. We showed that this measure can be used to generate substantially more accurate probabilistic forecasts than existing probabilistic forecast-aggregation models in the literature, including several recently proposed aggregation approaches (Kurvers et al., 2019; Palley & Soll, 2019; Prelec et al., 2017; Satopää et al., 2016), particularly when meta-probability weighting is combined with a simple recalibration approach. Furthermore, we showed that this measure was more effective at identifying expertise than other existing expertise-identification approaches in the single-question domain, including the weighting mechanism contained in the SP algorithm that we had identified in Chapter 2 and the recently-proposed Decision Similarity measure (Kurvers et al., 2019).

In Chapter 4, we compared the efficacy of the meta-probability weighting approach against cross-domain and within-domain weighting approaches, where forecasters' past performance may be known to the decision maker. We found that the meta-probability weighting approach performs poorly in comparison to these other aggregation approaches that rely on forecasters' past performance, which is perhaps unsurprising given that these other aggregation approaches make use of information about the expertise of forecasters contained in their responses to multiple questions. Nonetheless, the meta-probability weighting approach outperformed a cognitive modeling approach on a number of domains. Interestingly, we found that weighting forecasters by their performance on questions in unrelated domains was typically just as effective as weighting by their performance on questions in the same domain, suggesting that forecasters' expertise can be estimated with similar efficacy across many unrelated domains. Our results are consistent across a range theoretically unrelated domains and for both large and small crowds, and speaks to the generality of expertise found in crowds.

There are several useful applications of these findings in practice. In many real-world problems, decision makers rely on the information from groups of people in order to produce an informed decision. Individuals in the crowd can often differ in the level of expertise about a given problem, such that their performance might be consistently higher or lower than others in the crowd. By identifying and leveraging the views of individuals who are likely to generate accurate decisions, decision makers can improve the quality of collective decisions being made. Decision makers may also benefit from being able to distinguish between high-performing and low-performing individuals in the crowd, for example, for the purposes of recognising and rewarding high-performers or providing feedback and training to low-performers.

In problems where an individual's track record is well established, decision makers can typically distinguish between high-performing and low-performing individuals with high reliability. Unfortunately, in many applications of forecasting, the problem might be sufficiently novel that there are no records from which to identify the expertise of individuals. Measures of expertise that do not rely on these records of their past performance therefore have great practical utility. As we saw throughout this thesis, existing methods such as the Decision Similarity approach proposed by Kurvers et al. (2019) or weighting by forecaster confidence tend to be effective only under a narrow range of conditions and, in fact, appears to be generally less effective than the meta-probability weighting measure we developed. The current thesis thus provides a valuable contribution to the judgment and decision making literature in providing decision makers with better tools for identifying and leveraging the expertise of individuals in the crowd.

Despite the clear benefits of the meta-probability weighting approach developed in this thesis, there are a few important limitations that decision makers seeking to apply this approach should be aware of. As we demonstrated comprehensively throughout Chapter 3, the meta-probability weighting approach distinguishes between high-performing and low-performing individuals most effectively in crowds that contain a mixture of both groups and when the difference in expertise between groups is large. In problems where there is little-to-no systematic difference in expertise between individuals in the crowd, the meta-probability weighting approach is unlikely to offer meaningful improvements over basic aggregation approaches such as simple averaging. Decision makers may therefore wish to avoid expending additional resources for problems where they believe the crowd would be almost homogeneous in expertise, since the meta-probability algorithm

requires decision makers to elicit an additional set of meta-predictions for each decision.

While the meta-probability weighting algorithm is effective for identifying and leveraging latent expertise in the crowd, it is unable to account for the overlap of information between individuals in the crowd. As we showed in Chapter 3, the meta-probability weighting algorithm needs to be recalibrated, both in theory and in practice, due to (1) the overlap of information between forecasters, and (2) the bounded nature of the probability scale (Baron et al., 2014). We demonstrated one possible approach for combining the meta-probability weighting approach with other approaches that can account for these factors, such as a simple recalibration rule. A potential consideration for future research is to extend the theoretical model developed in Wilkening et al. (2020) that underpins meta-probability weighting to account for the sharing of information between individuals in the crowd, for example, by adapting the Pivoting approach developed by Palley and Soll (2019).

As misinformation and “fake news” becomes an increasingly greater concern in the modern socio-cultural landscape, the need for more refined approaches to distinguish between true experts and non-experts becomes more apparent. We hope that our research will, in part, inspire future researchers to find more innovative and efficient ways to deal with the challenges of harnessing the wisdom of crowds.

References

- Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners* (Vol. 30). Springer Science & Business Media.
- Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., . . . Mellers, B. (2016). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), 691–706.
- Baillon, A., Bleichrodt, H., Liu, N., & Wakker, P. P. (2016). Group decision rules and group rationality under risk. *Journal of Risk and Uncertainty*, 52(2), 99–116.
- Baillon, A., Tereick, B., & Wang, T. V. (2019). Follow the money: Bayesian markets to aggregate expert opinions when the majority can be wrong. In *Workshop on fintech and machine learning* (Vol. 5, p. 8).
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2), 133–145.
- Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 265–272.
- Blackwell, D., & Girshick, M. A. (1979). *Theory of games and statistical decisions*. Courier Corporation.
- Bolger, F., & Rowe, G. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, 35(1), 5–11.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly*

- Journal of the Royal Meteorological Society*, 135(643), 1512–1519.
- Bromme, R., Rambow, R., & Nückles, M. (2001). Expertise and estimating what other people know: The influence of professional experience and type of knowledge. *Journal of Experimental Psychology: Applied*, 7(4), 317.
- Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2), 267–280.
- Casati, B., Ross, G., & Stephenson, D. (2004). A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, 11(2), 141–154.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1.
- Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, 13(2), 128–152.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4), 559–583.
- Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2), 187–203.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Condorcet, M. d., Marie Jean Antoine Nicolas de Caritat. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. De l'Imprimerie royale.
- Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press on Demand.
- Cooke, R. M., & Goossens, L. L. (2008). Tu delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5), 657–674.
- Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? the surprising informational value of “just asking”. *Judgment and Decision Making*, 14(2), 135–147.
- Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise?

Decision, 1(2), 79.

- Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50), 15343–15347.
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69–106.
- Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, 57(6), 1082.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397), 171–185.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3), 519.
- Eteläpelto, A. (1993). Metacognition and the expertise of computer program comprehension. *Scandinavian Journal of Educational Research*, 37(3), 243–254.
- Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*, 1(2), 155–172.
- Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2), 137–146.
- Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7), 450–451.
- Genest, C., & McConway, K. J. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting*, 9(1), 53–73.
- Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1), 108–121.
- Gillen, B., McKenzie, J., & Plott, C. R. (2018). Two information aggregation mechanisms for predicting the opening weekend box office revenues of films: Boxoffice prophecy and guess of guesses. *Economic Theory*, 65(1), 25–54.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

- Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, 7(5), 398.
- Gronau, Q. F., & Wagenmakers, E.-J. (2019). Limitations of bayesian leave-one-out cross-validation for model selection. *Computational brain & behavior*, 2(1), 1–11.
- Hertwig, R. (2012). Tapping into the wisdom of the crowd-with confidence. *Science*, 336(6079), 303–304.
- 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.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430), 773–795.
- Knight, H. C. (1921). *A comparison of the reliability of group and individual judgments* (Unpublished doctoral dissertation). Columbia University (1921?).
- Koriat, A. (2008). Subjective confidence in one’s answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 945.
- Koriat, A. (2012). When are two heads better than one and why? *Science*, 336(6079), 360–362.
- Krueger, J. (1998). Enhancement bias in descriptions of self and others. *Personality and Social Psychology Bulletin*, 24(5), 505–516.
- 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.
- Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., ... Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, 5(11), eaaw9011.
- Lam, L. T., & Kirby, S. L. (2002). Is emotional intelligence an advantage? an exploration of the impact of emotional and general intelligence on individual performance. *Journal of Social Psychology*, 142(1), 133–143.
- 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., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9(3), 259.

- 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), 322–333.
- Lee, M. D., Steyvers, M., De Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science*, 4(1), 151–163.
- Lee, M. D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people’s rankings. *PloS one*, 9(5), e96431.
- Lovall, D., & Kahneman, D. (2003). Delusions of success. *Harvard Business Review*, 81(7), 56–63.
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion*, 8(4), 540.
- Mannes, A. E., Larrick, R. P., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology. social judgment and decision making* (pp. 227–242). New York, NY, US: Psychology Press.
- Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, 107(2), 276.
- Marschak, J., & Miyasawa, K. (1968). Economic comparability of information systems. *International Economic Review*, 9(2), 137–174.
- Marschak, J., & Radner, R. (1972). *Economic theory of teams (cowles foundation monograph 22)*. Yale University Press, New Haven, CT.
- Martinie, M., Wilkening, T., & Howe, P. D. (2020). Using meta-predictions to identify experts in the crowd when past performance is unknown. *Plos one*, 15(4), e0232058.
- Mayer, J., Salovey, P., Caruso, D., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion (Washington, DC)*, 1(3), 232.
- McCoy, J., & Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint arXiv:1703.04778*.
- McCoy, J. P. (2018). *Extracting more wisdom from the crowd* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Mellers, B., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? a multi-task, multi-benchmark study. *Judgment and Decision Making*, 12(4), 369–382.

- Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., . . . Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3), 267–281.
- Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision*, 3(1), 1.
- Meyer, A. N., Longhurst, C. A., & Singh, H. (2016). Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of crowdmed. *Journal of Medical Internet Research*, 18(1).
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, 23(1), 103–123.
- Müller-Trede, J., Choshen-Hillel, S., Barneron, M., & Yaniv, I. (2017). The wisdom of crowds in matters of taste. *Management Science*, 64(4), 1779–1803.
- Murphy, A. H., & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34, 273–286.
- Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature. *National Weather Digest*, 2(2), 2–9.
- Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7), 1330–1338.
- Murr, A. E. (2015). The wisdom of crowds: Applying condorcet’s jury theorem to forecasting us presidential elections. *International Journal of Forecasting*, 31(3), 916–929.
- Palley, A. B., & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, 65(5), 2291–2309.
- Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532.
- Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 71–91.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal*

- of Forecasting*, 30(2), 344–356.
- Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, 111(516), 1623–1633.
- Shlomi, Y., & Wallsten, T. S. (2010). Subjective recalibration of advisors’ probability estimates. *Psychonomic Bulletin & Review*, 17(4), 492–498.
- 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., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others’ opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 780.
- Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of experimental Psychology*, 15(5), 550.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Tetlock, P. E. (2017). *Expert political judgment: How good is it? how can we know?* Princeton University Press.
- Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.
- Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine learning*, 95(3), 261–289.
- Wagenmakers, E.-J., Lee, M., Rouder, J. N., & Morey, R. D. (2019). The principle of predictive irrelevance, or why intervals should not be used for model comparison featuring a point null hypothesis. *PsyArXiv*.
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... others (2018). Bayesian inference for psychology. part ii: Example applications with jasp. *Psychonomic bulletin & review*, 25(1), 58–76.
- Wilkening, T., Martinie, M., & Howe, P. (2020). *Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems*. (Manuscript submitted for publication)

- Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting*, 5(4), 605–609.
- Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of economic perspectives*, 18(2), 107–126.
- Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260–281.
- Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6, 1.

6 | Appendices

6.1 Chapter 2 Appendices

6.1.1 Manuscript detailing the theoretical model proposed in Chapter 2

This manuscript adapted some of the results from Chapter 2. Note that this manuscript was authored primarily by Tom Wilkening, who developed the theoretical model discussed throughout Chapters 2 and 3.

Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems

(Authors' names blinded for peer review)

Modern forecasting algorithms use the Wisdom of Crowds to produce forecasts better than those of the best identifiable expert. However, these algorithms may be inaccurate when crowds are systematically biased or when expertise varies substantially across forecasters. Recent work by Prelec et al. (2017) has shown that meta-predictions—a forecast of the average forecast of others—can be used to correct for biases even when no external information such as forecasters' past performance is available. We explore whether meta-predictions can also be used to improve forecasts by identifying and leveraging the expertise of forecasters. We develop a confidence-based version of the Surprisingly Popular algorithm of Prelec et al. (2017). Like the original algorithm, our new algorithm is robust to bias. However, unlike the original algorithm, our version is predicted to always weight forecasters with more informative private signals more than forecasters with less informative ones. In a series of experiments, we find that the modified algorithm does a better job in weighting informed forecasters than the original algorithm and show that individuals who are correct more often on similar decision problems contribute more to the final decision than other forecasters. Empirically, the modified algorithm outperforms the original algorithm for a set of 500 decision problems.

Key words: expertise, meta-knowledge, wisdom of crowds, forecasting, aggregation

1. Introduction

The Wisdom of Crowds has revolutionized the way in which we make predictions. It is the phenomenon where crowds make consistently better predictions, judgments, or estimates than even the most-expert individuals (Galton 1907, Surowiecki 2005). The superiority of aggregate predictions over individual predictions has been demonstrated across a variety of domains, but has gained particular attention in economic, political, and market forecasting where there are often high stakes involved (Budescu and Chen 2015, Dreber et al. 2015, Mellers et al. 2015, Müller-Trede et al. 2017, Tetlock 2017, Gillen et al. 2018).

The simplest approach to aggregating predictions is to use majority voting. As shown by the Condorcet Jury Theorem (Condorcet 1785), the probability that majority voting produces the correct decision for a binary decision increases towards 100% as the group sizes increases, under the assumption that each individual is more likely to be correct than incorrect. Despite its appealing properties, majority voting may often be inaccurate when crowds contain a large proportion of

uninformed forecasters or when the population of forecasters are systematically biased (Simmons et al. 2011, Budescu and Chen 2015).

To deal with the issue of uninformed forecasters, researchers have developed aggregation techniques that use training data to identify and weight forecasters based on their expertise. For example, Cooke (1991) developed a model that identifies and excludes non-experts from the crowd based on their performance on seed questions with outcomes that are known to the decision-maker. Similarly, Budescu and Chen (2015) showed that significant improvements in accuracy over the unweighted mean could be obtained by weighting experts by their performance relative to the crowd and excluding forecasters who did not improve the aggregate prediction.

Although expert-selection methods often generate better predictions than majority voting, researchers are not always able to identify individuals with the relevant expertise in advance. For example, forecasters' performance on prior problems with known outcomes might not effectively predict performance on problems of actual interest, and collecting the responses to a panel of relevant problems may be impractical (Genre et al. 2013, Clemen 1989). We refer to forecasting problems where it is either not possible or not helpful to use the individual's responses to prior problems as "single-question" forecasting problems, as the task is then to make the best forecast possible based on data relating only to a single question. We concentrate on the single-question problem for the rest of the paper.

The standard approach to the single-question prediction problem has been to use reported confidence to weight forecasters or to simply select the answer with the highest confidence (Koriat 2012, Prelec et al. 2017). These confidence-based approaches treat confidence as a predictor of expertise, weighting more-confident judgments more than less-confident judgments in the aggregation process. However, forecasters who hold the majority opinion tend to be overconfident while individuals who hold the minority opinion tend to be under-confident (Hertwig 2012, Koriat 2008, 2012). Thus, confidence may be negatively correlated with accuracy in "wicked" problems where most forecasters are incorrect. Indeed, there are many examples in the literature in which incorrect forecasters are more confident in their forecasts than correct forecasters (Koriat 2008, 2012, Fischhoff and MacGregor 1982, Lee and Lee 2017).

In this paper we explore whether we can improve upon existing confidence-based approaches by combining forecasts with meta-predictions about the forecasts of others. In a remarkable paper, Prelec et al. (PSM, 2017) proposed a novel algorithm that uses meta-predictions to correct for crowd biases. Their Surprisingly Popular (SP) algorithm generates predictions by using forecasters' votes about whether a particular event will be true or false and forecasters' meta-predictions—a prediction of the proportion of other forecasters that will vote true. The SP algorithm predicts the outcome that is more popular than the crowd expects (i.e., the surprisingly popular outcome) to

be the correct answer. In other words, the SP algorithm predicts true when the total number of true votes exceeds the average of the meta-predictions, and false otherwise.

PSM showed that the SP algorithm has the important theoretical property that it will always predict the correct answer when aggregating reports from a large homogeneous population of Bayesian forecasters, even when a substantial fraction of these forecasters are biased. In the first section of this paper, we show that an alternative Surprisingly Confident (SC) algorithm, which generates predictions by using forecasters' confidences and meta-predictions about the confidences of others, also shares this property. We then explore the theoretical properties of the SP and SC algorithms as they relate to expertise.

In our theoretical framework, we consider an environment in which individuals are asked binary true or false problems and share a common prior about the likelihood that the answer is true. Individuals receive signals from an information system and form a posterior belief about whether the answer is true using Bayes rule. The posterior belief held by an individual influences both their vote and their meta-prediction of the votes of others. We say that an individual has *a more informative private signal* than another individual if (i) the two individuals have posteriors that are both above the common prior or below the common prior and (ii) the absolute distance between the first individual's posterior and the common prior is larger than the second. An algorithm leverages informed forecasters if individuals with more informative signals contribute more to the algorithm's final prediction than those who are uninformed.

Our first result is that the SP algorithm actually leverages *uninformed* forecasters in problems where the crowd is initially unbiased. That is, the contribution that an individual makes to the final prediction of the algorithm is decreasing in the quality of the individual's information, such that individual forecasters who receive the most information have lower contributions to the aggregated forecast than individual forecasters who receive less information. To prove this result, we provide a novel alternative formulation of the SP algorithm, which expresses the algorithm in terms of a weighted average of the forecasters' votes. In this formulation, the weight assigned to an individual's vote is proportional to the absolute difference between a forecasters vote and their meta-prediction about the vote of others. We show that in unbiased problems, the weights are largest for fully uninformed individuals and strictly decrease as an individual becomes better informed.¹

Our modified SC algorithm improves on the way the algorithm weights forecasters with better-informed signals. Specifically, the SC algorithm always leverages forecasters with more informative

¹ This result does not mean that the SP algorithm will always perform poorly in unbiased settings. We show in our examples that the SP algorithm reveals important information that is common knowledge to all forecasters regarding the structure of signals. In large samples, this information is enough to correctly predict the right answer in both biased and unbiased problems when all forecasters know the true distribution of potential signals.

private signals regardless of whether the decision problem is biased or unbiased. Thus, a forecaster who has a more informative signal will always make a larger contribution to the final outcome of the algorithm than one who has a less informative signal. We show that the differences in the weight functions between the SP and SC algorithm may lead the SC algorithm to be more accurate than the SP algorithm when the sample of forecasters is finite, particularly in cases where forecasters vary in expertise.

Although our first result suggests that the SP algorithm may over-weight uninformed individuals and under-weight informed ones, it ignores a key advantage of crowd forecasts. In problems with an unbiased prior, the votes of forecasters who receive no information will be random while the votes of those who know the correct state will be perfectly correlated. This will cause the votes of the uninformed forecasters to partially cancel out as crowd size increases and may offset the weighting of individuals.

To understand the aggregate properties of the SP and SC algorithms, we consider a more general environment in which individuals share the same prior belief but have access to one of two information systems that are ordered in terms of informativeness. We refer to individuals who draw signals from the more informative information system as experts and individuals who draw signals from the less informative system as novices. Although experts and novices are assumed to have the same priors, experts are expected, on average, to receive more informative private signals and therefore predict the correct answer more often than novices. An algorithm leverages this expertise if the expected contribution of an expert is greater than that of a novice in both true and false questions.

As a second result, we show that the SC algorithm will leverage expertise in any environment where private signals are independent after conditioning on the state.² By contrast, the SP algorithm requires additional assumptions to ensure that the algorithm leverages expertise. In Appendix B, we derive a set of sufficient conditions on the structure of the information systems that guarantee that the SP algorithm leverages expertise. Our conditions suggest that in unbiased problems, experts will be leveraged by the SP algorithm in environments where (i) there is a mix of both experts and novices in the population and (ii) novices are reasonably uninformed.

Finally, we consider the properties of the SC algorithm in more realistic settings where reported confidences do not coincide with each forecaster's posterior and where forecasts are systematically miscalibrated. We show that even when forecasters are not Bayesian, the SC algorithm will predict

² The SC algorithm is able to leverage expertise in cases where individuals share a biased common prior. Such priors may come from a commonly observable public signal. Thus the algorithm will also leverage expertise in environments where individuals receive both a commonly observed public signal and conditionally independent private signals. See Palley and Soll (2018) for an alternative probabilistic forecasting algorithm that is designed to account for more complex signal correlation structures.

the true answer in large samples if (i) reported confidences are weakly increasing in the underlying true posteriors and (ii) forecasters take systematic overconfidence into account when reporting their meta-prediction. This result suggests that the algorithm is likely to perform well in settings where forecasters who believe the consensus position is correct are overconfident and forecasters who believe the consensus position is incorrect are under-confident. In such environments, other confidence-based aggregation approaches tend to fail.

Our theoretical results predicts particular patterns in the weights generated in the SP algorithm that vary with initial crowd bias. To analyse whether these patterns exist empirically, we estimate the relationship between weights and signals in two datasets: a replication of the US States dataset of PSM in which the prior is predicted to be strongly biased, and a new quiz dataset where we can vary the distribution of experts and novices by varying task difficulty. Using the probabilistic forecasts of an individual as a proxy for their posterior belief, we show that the weights in the datasets from both our experiments follow the patterns predicted by the theory for both the SP algorithm and the SC algorithm.

Our theoretical model also predicts that in unbiased problems, the SP algorithm is likely to perform well when there is variation in experts and non-experts in the environment. To test for this feature, we systematically vary the difficulty of problems in our new quiz dataset to create variation in problem difficulty and the likely mix of novices and experts. Consistent with our theoretical predictions, the SC algorithm leveraged expertise more effectively than the SP algorithm for all difficulty levels.

Finally, we compare the performance of the SC and SP algorithms across our two experiments. We find that the SC algorithm outperforms the SP algorithm in our quiz datasets and that this outperformance is driven by difficult problems where the SC algorithm performs well. Surprisingly, the SC algorithm performs poorly in easy quiz problems. We discuss how this may be due to the treatment of commonly observed signals in the SC algorithm and briefly discuss how the SC and SP algorithms might be combined to improve forecasts over each algorithm on its own.

Our paper contributes to the literature by providing a single-question algorithm that has promising empirical and theoretical properties in terms of expertise. The SC algorithm is robust to bias and corrects for overconfidence in situations where other confidence-based aggregation approaches fail. Further, under reasonable assumptions, the SC algorithm has the intuitive feature that uninformed individuals will be given zero weight and maximally informed individuals will be given the highest weight.³

³ The weights used in the SC algorithm can also be used in probabilistic forecasting problems. See Martinie et al. (2019) for a discussion of how the weights of the SC algorithm can be adapted to the probabilistic forecasting domain and for a comparison of the algorithm to other probabilistic forecasting algorithms proposed by Palley and Soll (2018) and Satopää et al. (2016). See McCoy and Prelec (2017) and Palley and Satopää (2020) for two alternative approaches for using meta-predictions in the forecasting domain.

The rest of the paper is structured as follows. We present our main theoretical results in Section 2 and test these results empirically in Section 3. We collect all proofs for the lemmas and propositions in Appendix C.

2. Theory

We consider a Bayesian model in which a crowd of N forecasters is assembled to predict the outcome of a single event. The outcome of the event, $o \in \{T, F\}$, is binary and can be true or false. Forecasters share a common prior $p(T)$ that the event is true.

Each forecaster receives a private signal S , that is a random variable taking on real value realisations in the set $\{s_1, \dots, s_m\} \cup \{s_\emptyset\}$ where $0 \leq s_1 < s_2 < \dots < s_m \leq 1$ and $s_1 < s_\emptyset < s_m$. As our outcome space is binary, it is without loss of generality that we normalise the signals so that their value is equal to the posterior belief that an event is true. That is, $s_j := p(T|s_j)$. We let $s_\emptyset$ represent the case where an individual receives an uninformative signal so that $s_\emptyset := p(T)$.

To minimise ambiguity, we will use s_j to denote the j th lowest posterior in the set $\{s_1, \dots, s_m\} \cup \{s_\emptyset\}$. Thus, it is always the case that $s_1 < s_2$. We will use σ_k to denote the signal drawn by a particular forecaster k . As each signal is drawn randomly, there is no inherent order between σ_1 and σ_2 .

We use a left stochastic matrix called an *information service* to model the distribution of signals across forecasters in each state.⁴ Initially, we will assume that all participants receive signals from the same information system denoted as Q .⁵ We also assume that the properties of Q are common knowledge to all forecasters.

An information service is composed of a likelihood matrix $[Q_{oj}]_{2 \times (m+1)}$. Each element of the first row of Q represents the probability that the signal is s_j given the outcome is $o = T$. Likewise, each element of the second row of Q represents the probability that the signal is s_j given the outcome is $o = F$. For ease, we will denote the first row elements with T and the second row elements with F . Thus $Q_{Tj} := Q_{1j} = p(s_j|T)$ while $Q_{Fj} := Q_{2j} = p(s_j|F)$.

We note two important features of an information service. First, an information service acts as a transition matrix from a state of nature to a signal and thus $\sum_j Q_{oj} = 1$ for each row $o \in \{T, F\}$. Second, upon receiving a message from an information service, agents revise their priors using Bayes rule. For any signal that occurs with positive probability (i.e., where $Q_{Tj} + Q_{Fj} > 0$), the posterior belief that the event is true is given by

$$p(T|s_j) = \frac{p(T)Q_{Tj}}{p(T)Q_{Tj} + p(F)Q_{Fj}}.$$

⁴ See Blackwell (1953), Blackwell and Girshick (1979), Marschak and Miyasawa (1968), Marschak and Radner (1972) for general treatments of information systems.

⁵ In Subsection 2.3, we will relax this assumption and introduce experts who will receive signals from a more informative information service.

By construction, this is equal to s_j for all signals that occur with positive probability.

It will be useful to classify decision problems based on the properties of Q . The following definitions help to identify three types of decision problems, which will respond differently across aggregation problems. We first classify decision problems based on whether the common prior is biased or unbiased:

DEFINITION 1. A decision problem has an **unbiased prior** if $s_\emptyset = 0.5$ and a **biased prior** if $s_\emptyset \neq 0.5$.

We further divide the class of unbiased problems into asymmetric and symmetric decision problems. We will call an information service *symmetric* if the likelihood of posterior s_i in state T is equal to the likelihood of posterior $(1 - s_i)$ in state F . Symmetry places restrictions both on the set of outcomes and on the relationship between likelihoods.

DEFINITION 2. An information system is **symmetric** if (i) $s_\emptyset = \frac{1}{2}$, (ii) the cardinality of the set $\{s_1, \dots, s_m\}$ is even, and (iii) $Q_{Ti} = Q_{F(m-i+2)}$ for all $i \in \{1, \dots, m+1\}$.

Following Prelec et al. (2017), we will focus attention to information systems that have the following property:

DEFINITION 3. An information system Q is **responsive** if there is a positive probability that a forecaster votes for the correct answer both when the state is true and when the state is false:

$$\sum_{\{i|s_i \leq .5\}} Q_{Fi} > 0 \quad \text{and} \quad \sum_{\{i|s_i \geq .5\}} Q_{Ti} > 0$$

Responsive information systems require that the bias is not so strong that all forecasters will go against their own private information and vote with the publicly observable signal in large samples. The assumption will imply that the expected vote in the true state is larger than the expected vote in the false state.

Finally, we will use the following partial ordering of signals to evaluate how the algorithm treats individuals with different amounts of information.

DEFINITION 4. Forecaster i has a **more informative private signal** than forecaster j if either (i) $\sigma_i < \sigma_j < s_\emptyset$ or (ii) $\sigma_i > \sigma_j > s_\emptyset$.

Intuitively, the informativeness of a forecaster's private signal is related to the distance between his posterior and the common prior. We have restricted attention to cases where σ_i and σ_j are either both greater than $s_\emptyset$ or both less than $s_\emptyset$ so that distance is directly related to the relative changes in the likelihood ratios of the two forecasters.⁶

⁶ For example, if $s_i > s_j > s_\emptyset$, then $\frac{Q_{Ti}}{Q_{Fi}} > \frac{Q_{Tj}}{Q_{Fj}} > \frac{p(T)}{p(F)}$. Thus $|s_i - s_\emptyset| > |s_j - s_\emptyset|$ implies $|\frac{Q_{Ti}}{Q_{Fi}} - \frac{p(T)}{p(F)}| > |\frac{Q_{Tj}}{Q_{Fj}} - \frac{p(T)}{p(F)}|$.

We note that the ordering of private signals is related to the extremity of the posterior, but is not equivalent to extremity in decision problems where there is a biased priors. For example, in a problem where the common prior is $s_{\emptyset} = .75$, a forecaster who has a signal of $\sigma_i = 0.5$ will have received a more informative signal than a forecaster with a signal of $\sigma_j = 0.6$.

2.1. Single-question forecasting algorithms

We consider single-question forecasting algorithms that use information from predictions and meta-predictions about the current event only. Let $V_i(T|\sigma_i) \in \{0, 1\}$ be the forecaster's prediction, or vote, that the event is true given signal σ_i , and let $P_i(T|\sigma_i) \in [0, 1]$ be the forecaster's probabilistic forecast that the event is true. Further let $M_i^V(Q|\sigma_i) \in [0, 1]$ be a forecaster's **vote meta-prediction**: a forecaster's meta-prediction of the share of other forecaster's that will vote true. Let $M_i^P(Q|\sigma_i) \in [0, 1]$ be a forecaster's **probability meta-prediction**: the forecaster's meta-prediction of the average probability forecast of all other forecasters. To simplify notation, we let $V_i := V_i(T|\sigma_i)$, $P_i := P_i(T|\sigma_i)$, $M_i^V := M_i^V(Q|\sigma_i)$, and $M_i^P := M_i^P(Q|\sigma_i)$.

We let $X_i := (V_i, P_i, M_i^V, M_i^P)$ be forecasters i 's full report and let $X = (X_1, X_2, \dots, X_N)$ be the full reports of all forecasters. Each algorithm we consider is a mapping $T : X \rightarrow \{0, 1\}$, which aggregates the data from a single event into a categorical forecast of whether the event is true or false. We assume the forecasters are truthful in all the algorithms and that they randomise their votes uniformly if they have the *uninformed* posterior of 0.5. This implies that $V_i = 0$ if $P_i < 0.5$, $V_i = 1$ if $P_i > 0.5$, and V_i is equally likely to be zero or one when $P_i = 0.5$.

We explore the theoretical properties of two alternative meta-prediction algorithms in this paper: the Surprisingly Popular (SP) algorithm of Prelec et al. (2017) and a variant that we refer to as the Surprisingly Confident (SC) algorithm. In the SP algorithm, the proportion of the crowd voting true is compared to the mean vote meta-prediction. If the proportion of true votes exceeds the average of the vote meta-prediction, the event is predicted to be true. Otherwise, the event is predicted to be false. Formally,

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N (V_i - M_i^V) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Although the standard formulation of the SP algorithm is easy to compute, it is relatively difficult to understand how the algorithm treats forecasters with different signals. The following lemma provides an alternative “weighted average” formulation of the SP algorithm that helps to make clear how individuals with different information are treated in the algorithm. As seen in the proof located in Appendix C, the transformation from one formulation to the other is mechanical and does not rely on any assumptions regarding the signals received by forecasters and their votes or vote meta-predictions.

LEMMA 1. *The SP algorithm can be rewritten as*

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N W_i^{SP} V_i > 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where each forecaster's weight is given by the normalised absolute difference between that forecaster's vote and their vote meta-predictions:

$$W_i^{SP} := \frac{|V_i - M_i^V|}{\sum_{j=1}^N |V_j - M_j^V|}.$$

Proof: All proofs are collected in Appendix C.

In the weighted average formulation of the SP algorithm, the weights are constructed so that $\sum_{i=1}^N W_i^{SP} = 1$. Thus, the weight given to each individual forecaster is proportional to $|V_i - M_i^V|$, the absolute difference between the forecaster's vote and the forecaster's meta-prediction about the votes of others.

The alternative SC algorithm uses probabilities and probability meta-predictions to predict the true outcome. Analogous to the SP algorithm, the average probabilistic forecast (or confidence) is compared to the mean probability meta-prediction. If the mean probabilistic forecast is larger than the mean probability meta-prediction, the event is predicted to be true. Otherwise, the event is predicted to be false. Formally,

$$T_{SC}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N (P_i - M_i^P) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Like the SP algorithm, the SC algorithm can be represented as a weighted average. In this representation

$$T_{SC}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N W_i^{SC} \mathbb{I}_{\{P_i > M_i^P\}} > 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where (i) $\mathbb{I}_{\{P_i > M_i^P\}}$ is an indicator variable that is one when a forecaster's probability forecast exceeds their probability meta-prediction and zero otherwise and (ii) each forecaster's weight is given by the normalised absolute difference between the forecaster's probabilistic forecast and their probability meta-prediction:

$$W_i^{SC} := \frac{|P_i - M_i^P|}{\sum_{j=1}^N |P_j - M_j^P|}.$$

The weighted version of the SC algorithm has the same structure as the weighted version of the SP algorithm, but has two differences. First, the algorithm uses the difference between a forecaster's

probabilistic forecast (or confidence) and their probability meta-prediction (rather than their vote and vote meta-prediction) to identify whether a forecaster should be recorded as a zero or a one in the final aggregation. As discussed below, an individual who receives $\sigma_i > s_\emptyset$ is predicted to have a probability forecast that exceeds their probability meta-prediction while the opposite is true when $\sigma_i < s_\emptyset$. Thus, the algorithm assigns a forecaster the equivalent of a true vote when they have a signal greater than the prior, and a false vote when they have a signal less than the prior. Second, the SC uses the probability forecasts and probability meta-predictions to generate the weights rather than using the votes to generate the weights. As discussed below this seemingly small adjustment has important implications in the way that the two algorithms weight forecasters with different signals.

2.2. Weights and Information

We first ask how the weights used in the SP and SC algorithms relate to information when all forecasters reports are consistent with Bayes rule. Intuitively, an algorithm will be able to best exploit the private information of forecasters if forecasters with more informative private signals contribute more to the algorithms final performance than those who have less informative private signals. The following propositions shows that the opposite relationship holds in the SP algorithm in situations where the prior is unbiased:

PROPOSITION 1. *In the SP algorithm, if (i) forecaster i has a more informative private signal than j and (ii) the prior is unbiased, then the weight given to forecaster i will be strictly less than the weight given to forecaster j .*

The intuition for Proposition 1 can be seen in the left side of panel (a) of Figure 1, which plots out the vote function and a typical vote meta-prediction function over all possible posteriors in the case of a symmetric information system, which has an unbiased prior. As can be seen by looking at the vote function, individuals will vote $V_i = 0$ when $\sigma_i < 0.5$ and $V_i = 1$ when $\sigma_i > 0.5$. Thus, the vote is a step function that switches exactly at the unbiased prior.

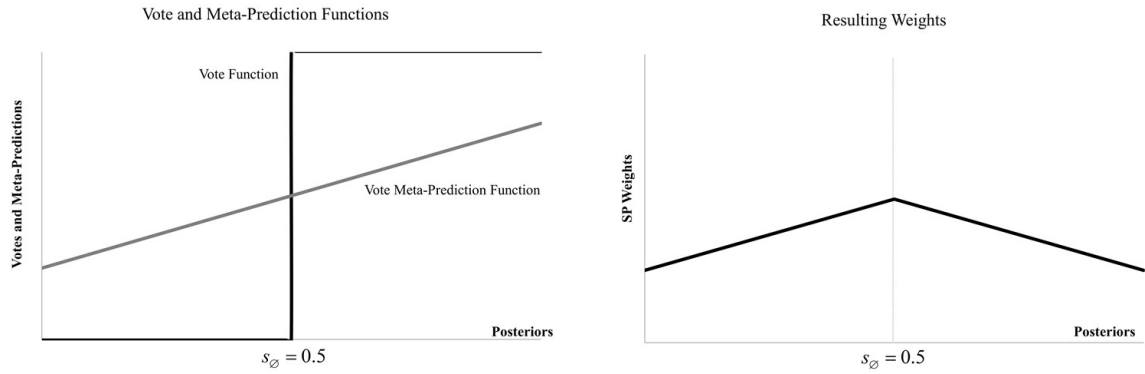
The vote meta-prediction of an individual is based on their belief about the votes made by all other participants. Given an outcome state o , the expected proportion of true votes from information service Q is given by

$$\mathbb{E}V(Q|o) = \sum_{\{i|s_i \geq .5\}} \gamma(Q_{oi}),$$

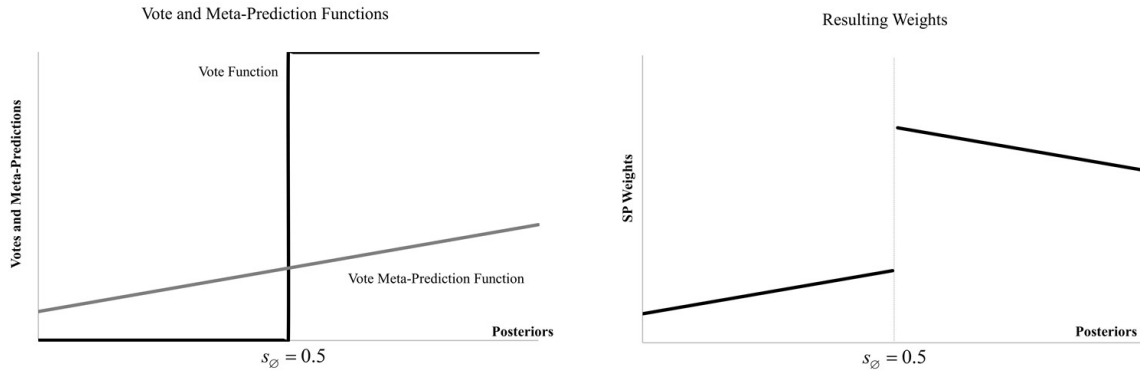
where $\gamma(Q_{oi}) = \frac{1}{2}Q_{oi}$ if $s_i = .5$ and $\gamma(Q_{oi}) = Q_{oi}$ otherwise. A forecaster with signal s_k 's vote meta-prediction about the average vote share from information service Q is

$$M^V(Q|s_k) = s_k \mathbb{E}V(Q|T) + (1 - s_k) \mathbb{E}V(Q|F).$$

(a) SP Weights in Symmetric Decision Problems



(b) SP Weights in an Asymmetric Decision Problem



(c) SP Weights in a Biased Decision Problem

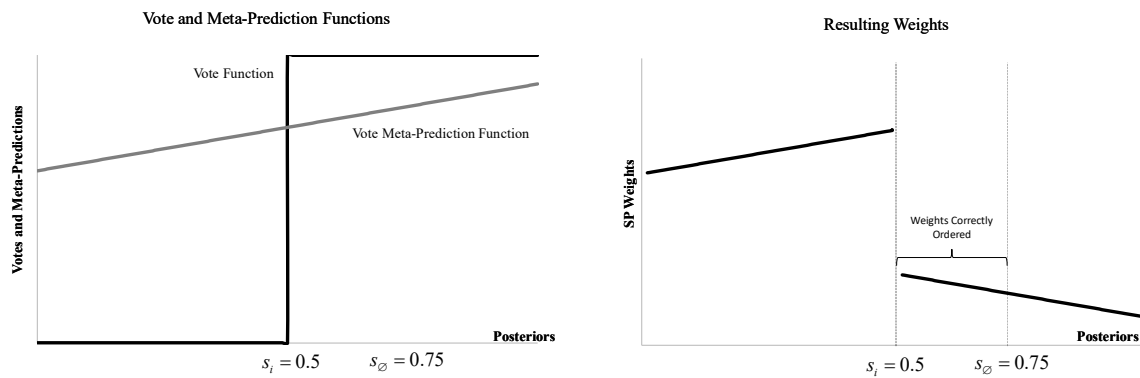


Figure 1 The left panels show the vote function and a typical vote meta-prediction function over all possible posteriors in (a) the case of a symmetric information system with an unbiased prior, (b) an asymmetric information system with an unbiased prior, and (c) a symmetric information system with a biased prior. The right panels show the weights assigned by the SP algorithm for each possible posterior in each of the three cases.

Noting that $\mathbb{E}V(Q|T) > \mathbb{E}V(Q|F)$ when the information system is responsive, $M^V(Q|s_k)$ is a linear function that is increasing in s_k . The underlying information system in panel (a) is symmetric, which implies that $M^V(Q|s_\emptyset) = 0.5$.

As seen in the right side of panel (a), the weights for each individual forecaster is equal to the absolute distance between the vote function and the vote meta-prediction function. This distance is decreasing as the forecasters signal moves away from the prior in both directions. Thus, individuals who have signals closer to the common prior will always have a larger weight than individuals who have signals that are farther away.

With a symmetric information system, the weighing function is also symmetric and all fully uninformed individual are equally weighted. This is not the case, however, when we consider asymmetric information systems. As seen in panel (b) of Figure 1, when the information system is asymmetric, $M^V(Q|s_\emptyset)$ does not necessarily pass through 0.5. As such, there is a gap in the weight function at $s_\emptyset$. This gap is the main way in which the SP algorithm is able to correct for asymmetries in the information system that leads majority voting algorithms to incorrectly predict the state. In particular, an individual who votes for true but predicts that others are more likely to vote false reveals commonly known information about the properties of the information system. This information is then used to increase the weights of individuals who vote against the most popular outcome.

Despite the algorithm taking advantage of information about the asymmetry of the information system, individuals who have signals closer to the common prior always have a larger weight than individuals who have signals that are farther away on the same side of the prior. This implies that forecasters with more informative private signals continue to receive smaller weights than comparable forecasters with less informative private signals.

Finally, when the prior is biased, if (i) forecaster i has a more informative private signal than j and (ii) both signals are between the biased prior of $s_\emptyset$ and the uninformed prior of 0.5, then i will be weighted more than j . However, this relationship is reversed in other parts of the distribution. This can be seen in the example shown in panel (c) of Figure 1, where forecasters have a prior of 0.75 and where weights are decreasing for posteriors greater than $s_\emptyset = 0.75$ and for signals that are below 0.5.

We now show that the weights in the SC algorithm is well ordered when it comes to the information contained in the forecaster's private signals:

PROPOSITION 2. *In the SC algorithm, if forecaster i has a more informative private signal than forecaster j , then the weight given to forecaster i will be strictly greater than the weight given to forecaster j .*

The intuition for Proposition 2 can be seen in the left side of panel (a) of Figure 2, which plots out the probability forecast function and a typical probability meta-prediction function over all posteriors in the case of a symmetric information system. As can be seen, the probability forecast function is a linear line with a slope of 1. The probability meta-prediction function is also linear and is based on their belief about the probability of all other participants. Given an outcome state o , the expected average forecast from information service Q is given by

$$\mathbb{E}P(Q|o) = \sum_{s_i} s_i Q_{oi}.$$

A forecaster with signal s_k 's probability meta-prediction about the forecast of others is given by

$$M^P(Q|s_k) = s_k \mathbb{E}P(Q|T) + (1 - s_k) \mathbb{E}P(Q|F).$$

By the law of iterated expectations, $\mathbb{E}P(Q) = s_\emptyset \mathbb{E}P(Q|T) + (1 - s_\emptyset) \mathbb{E}P(Q|F)$. Thus, $\mathbb{E}P(Q|T) > \mathbb{E}P(Q|F)$ and $M^P(Q|s_k)$ is a linear function that is increasing in s_k with a slope less than 1. The law of iterated expectations also implies that the two lines will intersect at the prior of 0.5. The net difference between the two lines generates a “v” shape that correctly orders forecasters in terms of the informativeness of their signals.

Panels (b) and (c) of Figure 2 show that the mechanism also correctly weighs forecasters according to the informativeness of their signals in asymmetric problems and biased problems. As seen in panel (b), asymmetric information systems do not substantially change the way the algorithm operates since both probabilities and probability meta-predictions increase linearly in the posterior. As seen in panel (c), in a biased problem, the probability function and meta-probability line cross at the prior. Thus, individuals who receive no signal will still have zero weight.⁷

The different pattern of weights has implications for the accuracy of the SP and SC algorithms. The **expected weight assigned to true in the SP algorithm** as N grows large is

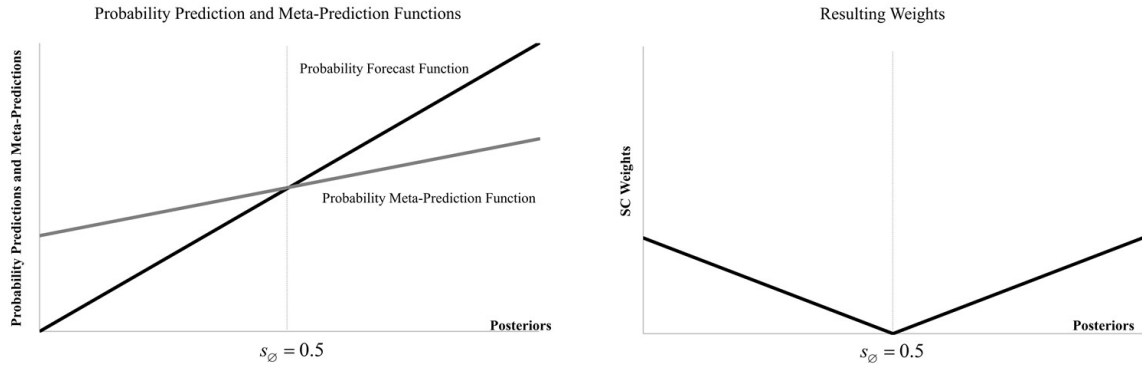
$$\mathbb{E}[W^{SP}] := \lim_{N \rightarrow \infty} \sum_i^N W_i^{SP} V_i.$$

Similarly, the **expected weight assigned to true in the SC algorithm** as N grows large is

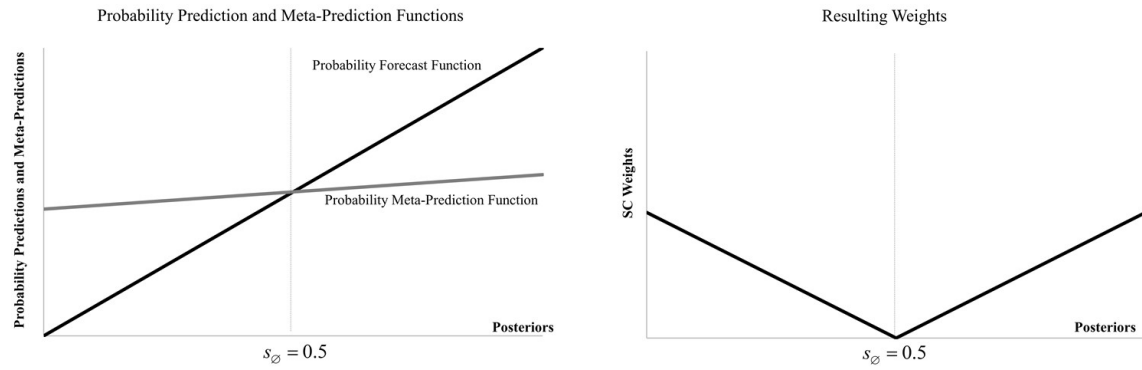
$$\mathbb{E}[W^{SC}] := \lim_{N \rightarrow \infty} \sum_i^N W_i^{SC} \mathbb{I}_{\{P_i > M_i^P\}}.$$

⁷ We note that the confidence-weighted algorithm, which calculates the average of all probabilistic forecasts and predicts true if this value is above 0.5 and false if the value is below 0.5, can also be written as a weighted average where the numerator of each weight is equal to $|P_i - 0.5|V_i$. This algorithm will generate “v” shaped weights that are centred at 0.5. Thus, in unbiased problems, if forecaster i has a more informative private signal than forecaster j , i will have a larger weight. This relationship does not hold, however, in biased problems because a forecaster with a more informative signal may have a posterior that is closer to 0.5.

(a) SC Weights in Symmetric Decision Problems



(b) SC Weights in an Asymmetric Decision Problem



(c) SC Weights in a Biased Decision Problem

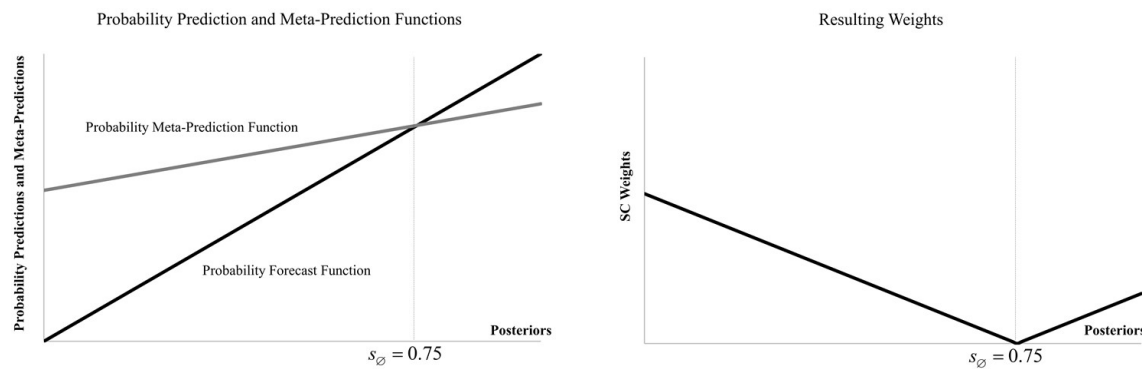


Figure 2 The left panels show a typical probability forecast function and probability meta-prediction function over all possible posteriors in (a) the case of a symmetric information system with an unbiased prior, (b) an asymmetric information system with an unbiased prior, and (c) a symmetric information system with a biased prior. The right panels show the weights assigned by the SC algorithm for each possible posterior in each of the three cases.

The following proposition shows that these expected weights are ordered in unbiased decision problems and that the SC algorithm will always assign more weight to the correct state as N grows large:

PROPOSITION 3. *For any unbiased information system, $\mathbb{E}[W^{SC}] \geq \mathbb{E}[W^{SP}]$ when the correct answer is true and $\mathbb{E}[W^{SC}] \leq \mathbb{E}[W^{SP}]$ when the correct answer is false.*

Based on the work by Prelec et al. (2017), $\mathbb{E}[W^{SP}] \geq 0.5$ when the correct answer is true and $\mathbb{E}[W^{SP}] \leq 0.5$ when the correct answer is false. Thus, in very large samples, both the SP and the SC answer will generate the correct answer. In small samples, the sample distribution will converge to a normal distribution with a mean equal to the expected weight assigned to true. This implies that if the variance of the two algorithms are the same, the SC algorithm will be more accurate than the SP algorithm.

In Appendix A, we report results from numerical simulations where we randomly constructed 100,000 unbiased information systems and calculated the variance in the total weight assigned to the correct state. We find that the variances of the two algorithms are similar in magnitude. We also calculate the sample size necessary to predict the correct state in 97.5% of cases under the assumption that the distribution is approximately normal in each sample. The SC algorithm requires a smaller sample size than the SP algorithm in over 99% of cases.

Appendix A also includes an analytic example where we explore how the SP and SC algorithms behave in a heterogeneous environment consisting of fully informed forecasters and forecasters who receive only weak signals. We show that in this setting, the SP algorithm may require much larger samples to ensure a high level of accuracy because the forecasters with weak signals will have large weights. This example suggests that the difference in weight functions may be important in difficult problems where there is only a small fraction of forecasters who know the correct answer. We study how the performance of the two algorithms relate to task difficulty in Section 3.

2.3. The Weighting of Experts

Although our first result suggests that the SP algorithm may over-weight uninformed individuals and under-weight informed ones, it ignores a key advantage of crowd forecasts. In problems with an unbiased prior, the votes of forecasters who receive no information will be random while the votes of those who know the correct state will be perfectly correlated. This will cause the votes of the uninformed forecasters to partially cancel out as crowd size increases and may offset the weighting of individuals.

To understand the aggregate properties of both algorithms, we consider a more general environment in which individuals have access to one of two information systems that are ordered in terms of informativeness. We refer to experts as individuals who draw signals from the more informative

information system and novices as individuals who receive draws from the less informative system. Thus an expert is defined as individual who is expected to be better informed about the correct answer prior to being asked a particular question.

We consider a variation of our baseline environment where we consider the limiting case where N is countably infinite. We divide forecasters in the population into two groups: experts and novices. Let Q^E be the information system used by expert forecasters and let Q^N be the information system used by novices. We assume that the proportion of experts in the crowd is known to all parties and given by $\theta \in [0, 1]$. We also assume that the properties of Q^E and Q^N are common knowledge.

We continue to assume that all forecasters make reports that are consistent with Bayes rule and we make three additional assumptions regarding the information services used by novices and experts.

ASSUMPTION 1. *Information service Q^E is more informative than information service Q^N : there exists a non-negative stochastic matrix $Z = [Z_{ki}]_{(m+1) \times (m+1)}$ such that*

$$Q^N = Q^E Z.$$

Assumption 1 says that when Q^E is more informative than Q^N , $Q_{oi}^N = \sum_k Q_{ok}^E Z_{ki}$. As we are multiplying across the rows of Q^E , we can interpret Z_{ki} as the conditional probability that when message k is received by Q^E , message i was received by Q^N . Thus $Z_{ki} = p(s_i | s_k)$ and Q^E is more informative than Q^N if it is possible to garble the signals of Q^E and generate Q^N . Note that Z is a non-negative stochastic matrix with $\sum_i Z_{ki} = 1$.

ASSUMPTION 2. *Experts and Novices draw conditionally independent signals: for a signal s_i from Q^N and a signal s_k from Q^E ,*

$$p(s_i, s_k) = p(s_i | T)p(s_k | T)p(T) + p(s_i | F)p(s_k | F)p(F).$$

ASSUMPTION 3. *Information system Q^E is responsive.*

Assumption 2 extends the assumption that signals are conditionally independent after conditioning on the state to an environment with two information services. The assumption rules out perverse situations where the garbling matrix creates additional information about the signals of others. Assumption 3 requires that at least expert forecasters will vote for the correct state with a positive probability. This assumption is necessary for the SP algorithm because it is vote based, but is not required for any result related to the SC algorithm.

Assumptions 1 and 2 imply that the information services are ranked but that signals from the two information services are independent once we condition for the state. Assumption 2 is sufficient for

the monotone likelihood ratio property (MLRP) to hold for signals between any two information services. This property implies that when an individual receives a high signal, he believes that other forecasters are also more likely to receive a high signal.

LEMMA 2. *For signals $s_i > s_j$ drawn from Q^t , $t \in \{N, E\}$, and signals $s_k > s_l$ drawn from Q^τ , $\tau \in \{N, E\}$, the monotone likelihood ratio property holds:*

$$p(s_i|s_k)p(s_j|s_l) > p(s_j|s_k)p(s_i|s_l). \quad (1)$$

Assumption 3 ensures that when the prior is biased, a subset of experts are willing to change their vote away from the prior for at least some realisation of the signal. Combined with MLRP, this assumption is enough to prove a modified version of PSM's theorem regarding the average estimates of the votes:

LEMMA 3. *In the SP algorithm, if Assumptions 1-3 hold, then the average estimate of the votes for the correct answer will underestimate the true proportion of votes for the correct answer as $N \rightarrow \infty$.*

The SP mechanism will predict the correct answer if the vote meta-prediction underestimates the true proportion of votes for the correct answer. Thus, Lemma 3 implies that the SP mechanism will continue to predict the correct answer in the limit when there are both experts and novices. The following lemma shows that the SC algorithm has a similar property when Assumption 2 holds:

LEMMA 4. *In the SC algorithm, if Assumptions 1-2 hold, then the average probability meta-prediction will be below the average probability forecast when the state is true and above the average probability forecast when the state is false as $N \rightarrow \infty$.*

Note that when $Q^E = Q^N = Q$, Assumptions 1 and 2 always hold. Thus Lemma 4 implies that the SC algorithm is robust to bias in the initial model where all forecasters draw signals from the same information system.

2.4. The Expected Total Contribution of an Expert or Novice

We now turn to the question of how the SP and SC algorithms weight experts and novices. Given information services Q^E and Q^N , a forecaster with signal s_k will make a vote meta-prediction of

$$M^V(\theta|s_k) := \theta M^V(Q^E|s_k) + (1 - \theta)M^V(Q^N|s_k),$$

where θ is the proportion of experts in the environment, $M^V(Q^E|s_k)$ is the vote meta-prediction of forecasters from information service Q^E , and $M^V(Q^N|s_k)$ is the vote meta-prediction of forecasters

from information system Q^N . Thus, the expected vote meta-prediction of forecasters in information service Q^t ($t \in \{N, E\}$) when the state is o is given by

$$\mathbb{E}[M^V(\theta|Q^t, o)] := \sum_k M^V(\theta|s_k)Q_{ok}^t.$$

Likewise, a forecaster with signal s_k will make a probability meta-prediction of

$$M^P(\theta|s_k) := \theta M^P(Q^E|s_k) + (1 - \theta)M^P(Q^N|s_k).$$

Thus, the expected probabilistic meta-prediction in information service Q^t ($t \in \{N, E\}$) when the state is o is given by

$$\mathbb{E}[M^P(\theta|Q^t, o)] := \sum_k M^P(\theta|s_k)Q_{ok}^t.$$

The quantities $[\mathbb{E}[V(Q^t|T)] - \mathbb{E}[M^V(\theta|Q^t, T)]]$ and $[\mathbb{E}[M^V(\theta|Q^t, F)] - \mathbb{E}[V(Q^t|F)]]$ represent the expected difference between the votes of forecasters with information service Q^t and their vote meta-prediction, for the true and false states respectively. We will call these quantities the **expected total contribution of an expert or novice** in the SP algorithm in state T and F respectively since they represent the total expected impact of a randomly selected individual from a given group taking into consideration both their vote and their vote meta-prediction under the given state. Similarly, we will call $[\mathbb{E}[P(Q^t|T)] - \mathbb{E}[M^P(\theta|Q^t, T)]]$ and $[\mathbb{E}[M^P(\theta|Q^t, F)] - \mathbb{E}[P(Q^t|F)]]$ the expected total contribution of an expert or novice in the SC algorithm.

In state T , the expected total contribution of an expert exceeds the expected contribution of a novice in the SP algorithm if

$$[\mathbb{E}[V(Q^E|T)] - \mathbb{E}[M^V(\theta|Q^E, T)]] > [\mathbb{E}[V(Q^N|T)] - \mathbb{E}[M^V(\theta|Q^N, T)]].$$

This leads us to our definition for leveraging expertise:

DEFINITION 5. An algorithm **leverages expertise** if the expected total contribution of an expert exceeds the expected contribution of a novice in all states.

In the SC algorithm, an individual's weight is strictly increasing in their signal. Using Blackwell's theorem, we can show the following result:

PROPOSITION 4. *The SC algorithm leverages expertise in all environments where Assumptions 1 and 2 hold.*

Proposition 4 shows that in a very general set of decision problems, the SC algorithm is able to leverage expertise. The result naturally generalizes to any number of information systems as

long as they are ranked in terms of informativeness. Thus, under a wide range of problems, the mechanism is predicted to leverage expertise.

In contrast, in Appendix B we provide two counter examples where the SP algorithm fails to leverage experts. These examples show that when the information systems is asymmetric, it is possible to find information systems where the total contribution of experts is less than that of novices in at least one state. Thus, Assumptions 1-3 are not sufficient to ensure that experts are leveraged and the SP algorithm may be less effective in heterogeneous environments. We provide two additional properties of the information system that are sufficient to ensure that the SP algorithm leverages expertise in symmetric decision problems and provide an example that helps to explain where these additional properties come from. The example suggests that the SP algorithm is likely to perform best in problems where there is a moderate number of experts.

2.5. Properties of the SC algorithm when confidence measures are noisy

Thus far we have considered how the SC algorithm behaves in an ideal setting where all forecasters are Bayesian and where the confidence elicited by each individual coincides with their posteriors. In this section we discuss some strengths and weaknesses of the SC mechanism that arise when we move from the this ideal setting to one where we incorporate the known biases that exist when eliciting confidences.

As noted in the introduction, a key issue for confidence-weighted algorithms is that they are sensitive to particular types of overconfidence. For instance, as discussed by Hertwig (2012), using confidences to weight forecasters can be problematic in environments where individuals who hold the majority opinion are overconfident while individuals who hold the minority opinions are underconfident. Such environments can arise when confidences are correlated with the majority viewpoint rather than perfectly relating to accuracy (Koriat 2008).

A surprising result is that in large samples, the SC algorithm will continue to correctly predict the correct state in settings where overconfidence occurs under two assumptions about confidences and probability meta-predictions. First, on average, confidences must be increasing in the underlying posterior of an individual forecaster. Second, forecasters must incorporate both their own overconfidence and the overconfidence of others into their meta prediction. We discuss this assumption below after a formal description of our result.

We begin by generalizing the model of Section 2.3 to allow for errors in the relationship between signals and reported confidences.

DEFINITION 6. Forecasters are **systematically miscalibrated** if there exists a weakly increasing function $c : [0, 1] \rightarrow [0, 1]$ and a right stochastic matrix $[R_{ij}]_{m+1, m+1}$ such that (i) the probability that an individual with posterior s_i reports confidence $c(s_j)$ is given by $R_{ij} := p(c(s_j)|s_i)$; (ii) for

any two posteriors $s_i > s_j$, $c(s_i) \geq c(s_j)$; (iii) there exists two posteriors $s_i > s_j$ that occur with positive probability where $c(s_i) > c(s_j)$; and (iv) $\sum_j c(s_j)R_{ij} = c(s_i)$ for all i .

Our definition of systematic overconfidence allows for forecasters to systematically misapply Bayes rule and to report confidences that are both too high and too low relative to the true posterior. The confidence function, $c(\cdot)$, allows for almost any non-decreasing mapping from true posteriors to confidence reports while the additional error structure allows for additional noise between signals and reports. This error structure is very general and can facilitate most behavioural patterns of overconfidence observed in the literature. In particular, it can accommodate the two main behavioural patterns of overconfidence discussed in Liberman and Tversky (1993) and Griffin and Brenner (2007): general overconfidence, the tendency for all forecasters to assign probabilities that are too close to 1 for the choice that they believe is correct; and specific overconfidence, the tendency for forecasters who believe one answer is correct to assign probabilities that are too close to 1 and for forecasters who believe the other answer is correct to assign probabilities that are too close to 0.5.⁸ It can also accommodate patterns of under-confidence, which is sometimes found in decision problems that are easy (Erev et al. 1994).

When forecasters are systematically miscalibrated, the average confidence of individuals from information service $t \in \{N, E\}$ in state $o \in \{T, F\}$ is given by

$$\mathbb{E}C(Q^t|o) = \sum_i \left(\sum_j c(s_j)R_{ij} \right) Q_{oi} = \sum_i c(s_i)Q_{oi}.$$

We will say that a forecaster's probability meta-prediction is **fully adaptive** if their meta-prediction (i) uses their confidence to assess the likelihood of each state of the world, and (ii) fully predicts the overconfidence of both novices and experts. Thus, an individual who is fully adaptive would report that the average confidence for forecasters from information service Q^t is:

$$M^C(Q^t|c(s_k)) = c(s_k)\mathbb{E}C(Q^t|T) + (1 - c(s_k))\mathbb{E}C(Q^t|F).$$

The following proposition shows that under the assumption of fully adaptive meta-predictions, the SC algorithm will generate the correct answer for any decision problem where confidence reports are systematically miscalibrated:

PROPOSITION 5. *If forecasters are systematically miscalibrated and all forecasters have fully adaptive meta-predictions, then the average probability meta-prediction will be below the average reported confidence when the state is true and above the average reported confidence when the state is false as $N \rightarrow \infty$.*

⁸ Although we allow only mean zero errors to be added at each confidence, the relatively weak conditions imposed on the confidence function, $c(\cdot)$, means that we can also model truncation bias that may occur when confidences have symmetric errors that are truncated on $[0, 1]$. In this case, $c(s_i)$ would simply be equal to the expectation of s_i over all realizations of the error.

Although our theoretical result requires a strong assumption about the average meta-prediction in the population, there are reasons to suspect that the algorithm will improve upon other confidence-weighted algorithms even when the assumption does not hold. As discussed in Koriatic (2008), Koriatic (2012), and Hertwig (2012), confidence-weighted algorithms typically fail in “wicked” problems where the position held by the consensus is wrong. In these problems, individuals who endorse the consensus answer tend to be over-confident while those who endorse the minority answer tend to be under-confident. For the SC algorithm to improve forecasts relative to the confidence-weighted algorithm, the average probability meta-prediction must be above 0.5 when the majority of forecasters vote ‘true’ but the correct answer is ‘false’, and below 0.5 if the majority of forecasters vote ‘false’ but the correct answer is ‘true’.⁹ This will be the case if the average probability meta-prediction and the consensus answer both lie on the same side of the uninformed prior. This relationship is likely to hold if beliefs about the consensus position not only influences each forecasters’ confidence report but also their belief about the confidence reports of others.¹⁰

We note that if forecasters reported $c(\sigma_i) = 0$ when $\sigma_i < 0.5$, $c(\sigma_i) = 1$ when $\sigma_i > 0.5$, and randomizes between 0 and 1 when $\sigma_i = 0.5$, then the fully adaptive meta-prediction would be to report the vote share. Thus, there exists a systematically miscalibrated decision problem where the reports of forecasters coincides with those elicited in the SP algorithm. This insight implies that in settings where forecasters are severely overconfident, the relative rankings of the two algorithms with respect to expertise and average vote weights may not hold. As such, we highlight some other strengths and weaknesses of the two algorithms before moving to the empirical section of the paper.

A clear advantage of the SP algorithm is that it elicits frequency information rather than probabilistic information from forecasters. Vote meta-predictions have the advantage that the forecasters do not have to estimate the level of overconfidence in the environment when forming their belief. Thus, vote meta-predictions may be more accurate in settings where overconfidence is present.¹¹ Further, a large literature exists that suggests that frequency information is encoded more naturally in the brain and may be more natural for individuals to express (Hintzman et al. 1982,

⁹ For example, suppose that the consensus answer is true, but the correct state is false. Then, if the average meta-prediction is 0.75, the confidence-weighted algorithm will correctly predict false if the average probability forecast is between $[0, 0.5]$ while the SC algorithm will correctly predict false if the average probability forecast is between $[0, 0.75]$.

¹⁰ Note that in cases where the consensus is correct, the SC algorithm will continue to predict the correct answer in large samples as long as the average probability meta-prediction is closer to the uninformed prior than the one calculated with forecasters who are fully adaptive. Thus, as long as forecasters don’t over predict the overconfidence of others, the SC algorithm and confidence-weighted algorithm are likely to both perform well in questions where the consensus is correct.

¹¹ Tereick (2019) argues that vote meta-predictions may be anchored towards the prior and proposes a self-aggregation algorithm that is more robust to these types of biases. Baillon et al. (2020) explores how to elicit incentive compatible meta-predictions using a market-based approach with randomized price offers.

Gigerenzer 1984, Gigerenzer et al. 1991). Thus, the SP algorithm is likely to have lower cognitive requirements than the SC algorithm.

Relative to the SP algorithm, the SC algorithm provides a larger communication space for providing information about their signals and meta-knowledge. In particular, the SC algorithm allows for forecasters to reveal that they are (i) uninformed or (ii) have limited insights into the information held by others and gives these forecasters little weight.

3. An Empirical Exploration of the SP and SC algorithms

In this section, we empirically estimate the weights generated in the SP and SC algorithms and study how these algorithms treat experts and novices. We concentrate our analysis on two experiments. The first is a replication of the US states capital dataset of Prelec et al. (2017). As seen in Prelec et al. (2017), forecasters in this dataset use what appears to be a heuristic based on population size to predict whether a city is a capital city in problems where they are uninformed. This heuristic naturally leads to a biased prior and is likely to lead to specific overconfidence — the tendency for forecasters who believe the consensus position is correct to be overconfident and forecasters who believe the consensus position is incorrect to be under-confident. We are interested in this environment since the SP is specifically designed to improve forecasting in biased environments and we would predict that this algorithm will perform well.

The second experiment uses a quiz dataset comprised of 500 problems that vary across five levels of difficulty. As seen in the theory section, the relative weighting of experts and novices is related to the proportion of experts in the environment. As we increase the difficulty of decision problems, we would expect the proportion of experts in the dataset to fall. We are thus interested in the relative performance of the SP and SC algorithm as we move from easy problems to hard ones, and we would predict that the SP algorithm leverages expertise most effectively with an intermediate number of experts.

We note that the actual expertise of individuals in our dataset is not observable and thus our empirical strategy requires us to proxy for expertise by using the track record of forecasters on other problems. This proxy is based on the assumption that expertise is correlated across questions and uses the fact that an individual who receives signals from a more informative information system will be correct more often than an individual who receives information from a less informative information system on average.

3.1. Experiment 1

Experiment 1 replicates PSM (2017)'s Study 1, which asked true or false questions about the capital cities of US states. For each state, participants were presented with the largest city and asked whether or not it was the state capital. This dataset provided a natural environment to study

the mechanisms underlying the SP algorithm's performance in a biased setting since PSM found in their original study that forecasters typically believe that the largest city in a state is the capital when they do not know the true answer. As this heuristic does not often predict whether a city is the state capital, the underlying information system is likely to be biased in favour of answering true. This allows informed individuals to make meta-predictions that differ substantially from their vote and potentially gives informed individuals large weights.

Our replication used a larger sample size than the original PSM study in order to compare the patterns of predictions and meta-predictions made by the best-performing and worst-performing forecasters in the crowd. In line with PSM, we collected forecasters' votes and meta-predictions about the average vote of others. Additionally, in order to compare the responses used by the SP algorithm and the SC algorithm, we also collected each forecaster's forecast of the likelihood that the event is true and their meta-prediction of the average forecast of all other forecasters.

3.1.1. Methods. We conducted the experiment online, with all participants recruited using Amazon Mechanical Turk. In PSM's experiments, forecasters were monetarily incentivised for accurately predicting the outcome as well as accurately predicting the proportion of the crowd endorsing each response. As our experiment was performed online, we removed the financial incentives to reduce the likelihood of participants looking-up the answer. We tested 100 respondents and only respondents inside the US were able to participate. Each survey was administered using Qualtrics, and participants were paid a flat fee of US \$2.50 for completing the survey. Participants were asked to answer each question as honestly as they could, and were asked not to cheat (e.g., by looking up any of the problems online). Eleven individuals who reported cheating at the task or had failed to complete the survey were excluded from the analyses, but were still paid. We completed data collection in January 2020 and analyses were conducted on the data of the remaining 89 participants.

The survey consisted of 50 trials (one for each US state, in alphabetical order of state). On each trial, participants were shown the sentence "X is the capital of Y" where X was the most populous city in the state Y. For example, on the first trial, all participants saw the bolded statement "Birmingham is the capital of Alabama." For each statement, participants were asked to answer four questions:

1. Is this statement more likely to be true or false?
2. What percentage of other people do you think thought the bolded statement was true?
3. What is the probability that the statement is true?
4. What is the average probability estimated by the other forecasters?

Forecasters were restricted to probabilities between 0 and 50 on question 3 if they reported that the statement was more likely to be false and between 50 to 100 if they reported that the statement was more likely to be true. Thus all participants were required to provide votes and probability forecasts that were consistent.

3.1.2. Weights in the SP and SC algorithms: Our theoretical model predicts that the SP weights assigned to individuals will decrease linearly as one moves away from the uninformative posterior of 0.5. However, because the states dataset is predicted to have a biased prior, we would predict that there will be a gap in the weight function at 0.5 and that this gap may lead false votes to be weighted more than true votes. To test for this, we use an individual's probabilistic forecast as a proxy for the forecaster's posterior¹² and estimate a linear weight function of the form

$$W_{ik}^{SP} = \alpha + \beta_1 |P_{ik} - 0.5| + \beta_2 V_{ik} + \epsilon_{ik}, \quad (2)$$

in which W_{ik} is the numerator of the SP weight of subject i in decision problem k , V_{ik} is their vote, P_{ik} is the probabilistic forecast and ϵ_{ik} are errors that are clustered at the individual and event level. We use the numerators of the SP weights here as they always fall between 0 and 1 and are fully comparable across problems. We predict that $\beta_1 < 0$, which would indicate that the weights are decreasing in the informativeness of the forecaster's signal between 0 and 0.5 and between 0.5 and 1. Based on PSM, we would also predict that $\beta_2 < 0$, which would indicate that the prior is biased towards true (see panel (c) in Figure 1 for the intuition).

For the SC algorithm, our theoretical model predicts that weights are upward sloping as one moves away from the prior.¹³ A proxy for this (unobserved) prior is given by the intersect between the identity line where the probability forecast is equal to itself and a regression line of the probability meta-prediction on the probability forecast. In the states data, this point is at 0.74. We then estimate a linear regression of the form

$$W_{ik}^{SC} = \alpha + \beta_1 |P_{ik} - 0.74| + \epsilon_{ik},$$

in which W_{ik}^{SC} is the numerator of the SC weight of subject i in decision problem k , P_{ik} is the probabilistic forecast and ϵ_{ik} are errors that are clustered at the individual and event level. We predict that $\beta_1 > 0$, which would indicate that the weights are increasing in the informativeness of signals. We find the following:

¹² In a Bayesian framework, an individual's forecast should be their posterior. Although this is not always the case empirically, probabilistic forecasts are strongly predictive of an individuals actual likelihood of being correct in the states dataset. Using a simple linear regression where we regress the probability of being correct on the absolute difference between an individual's probabilistic forecast and the uninformed prior of 0.5, an individual with a probabilistic forecast of 0.5 is correct 46.7 percent of the time while individuals with a probabilistic forecast of either 0 or 1 are correct 65.1 percent of the time.

¹³ Note that a biased prior therefore has a qualitatively different effect on the weighting function of the SC algorithm. In the SP weights, a biased prior leads to a gap in the weighting function at 0.5. In the SC weights, a biased prior leads to a shift in the kink point where forecasters are assigned the lowest weight.

Result 1 *Consistent with the theoretical model predictions, weights in the SP algorithm are decreasing in the distance from the 0.5 and there is a large gap in the weight function at 0.5. This gap leads to larger weights for false votes than for true votes. Weights in the SC algorithm are increasing in the distance away from the uninformed prior.*

Support for Result 1 is given in Figure 3, which plots the relationship between weights and the forecaster's posterior for the SP algorithm (top) and the SC algorithm (bottom). The black solid line in each graph is the predictions from the theoretical models while the dashed line is the estimates from a non-parametric kernel regression.

As seen in the top graph, the magnitude of forecasters' signals ($|P_{i,k} - 0.5|$) is a significant negative predictor of the forecasters' weight in the SP algorithm, $\beta_1 = -0.41$, $F(1, 88) = 40.61$, $p < .001$. Thus, consistent with our predictions, the SP weights appear to be decreasing in the distance away 0.5. Additionally, a forecasters' vote (V_{ik}) is a significant negative predictor of the forecasters weight, $\beta_2 = -0.24$, $F(1, 88) = 73.47$, $p < .001$. This can be seen by the apparent gap in the weight function at 0.5, which suggests a strong bias toward true responses in the dataset.¹⁴ The gap is large enough that the predicted weights of all forecasters voting false are larger than the weights of forecasters voting true in the model specification. As seen below, the gap helps the SP algorithm to predict the correct answer in most of the decision problems.

As seen in the bottom panel, the SC algorithm has weights that are increasing in the distance away from the predicted prior, with a significant and large positive slope in our model that is consistent with our predictions, $\beta_1 = .53$, $F(1, 88) = 69.5$, $p < .001$. On average, better-informed forecasters therefore are generating larger weights. The weights assigned to forecasters who predict that an event is false with certainty are particularly high, with an average weight that is at least twice as large as the weight assigned to any forecaster who voted true.

3.1.3. Expertise in the SP and SC algorithms: Having seen that the weights of our two algorithms match our theoretical predictions, we now explore how forecasters' total contributions relate to expertise. As a first approach, we ranked and sorted forecasters based on their mean accuracy computed using leave-one-out cross-validation and performed a median split between the best-performing individuals ("high-performers") and worst-performing individuals ("low-performers"). For the SP algorithm, we then compared the mean vote for each group to their mean vote meta-predictions for true and false problems separately. For the SC algorithm, we instead compared the mean probability forecast for each group to their mean probability meta-prediction.

¹⁴ As we show further below, weights in the SC algorithm are increasing in distance away from an uninformed prior of approximately 0.74. This implies that the gap in the SP weighting function is most likely due to a biased prior, rather than forecasters having access to asymmetrical information services. As we see in Appendix D, this also appears to be the case in Experiment 2.

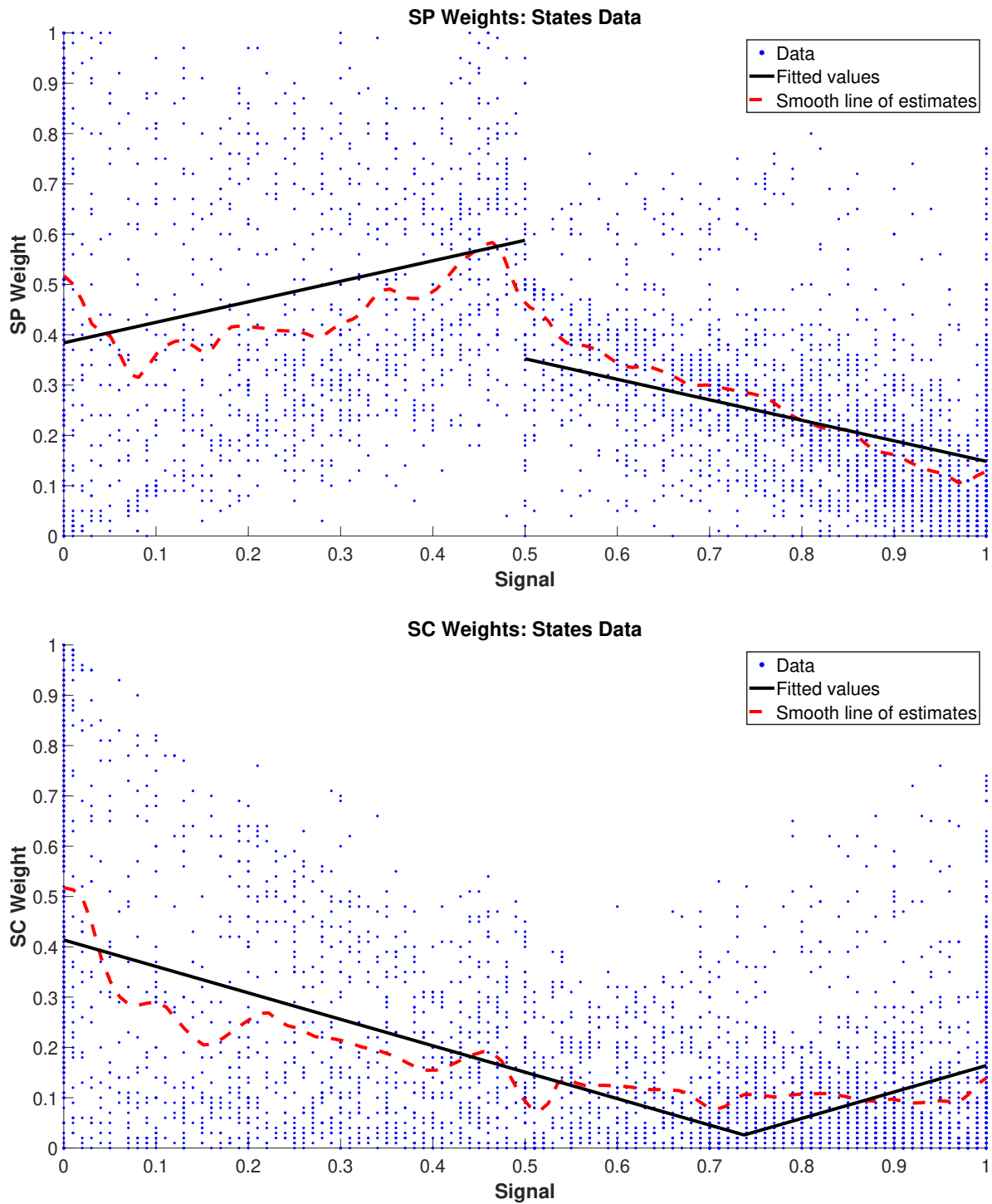


Figure 3 The relationship between forecasters' posterior and the weight assigned to them by the SP algorithm (top panel) and the SC algorithm (bottom panel) for the States Data. The solid black lines are the predictions from the theoretical models. The dashed line is from a non-parametric kernel regression.

The SP and SC algorithms leverage expertise if the average total contribution of an expert exceeds the average total contribution of a novice for both true and false problems. We find the following:

Result 2 *In the states data, the average total contribution of high-performers in the SP algorithm is statistically greater than that of low-performers in problems that are false, but there is no significant difference in problems that are true. The total contribution of high-performers in the SC algorithm is statistically greater than that of low-performers for both true and false problems.*

Support for Result 2 is provided in Figure 4, which compares the pattern of responses for high-performers and low-performers in the States dataset for both algorithms. The mean of high-performers' responses are shown as red circles, the mean of low-performers' responses are shown as blue crosses, and the shaded regions in these plots indicate where each algorithm would produce correct predictions and where the total contribution of the group has the correct sign. The horizontal (and vertical) distance from the reference line to each point corresponds to the absolute difference between that group's mean vote (or probability forecast) and their mean vote (or probability) meta-prediction. In the top panels, the distance between each point and the dotted line is therefore proportional to the total contribution to the SP algorithm for that particular group and event. Similarly, the distances in the bottom panels are proportional to the total contribution to the SC algorithm for each group and event.

We used paired sample t -tests to compare high-performers' and low-performers' average total contributions separately for problems where the outcome was true and problems where the outcome was false. As seen in the top left panel of Figure 4, high- and low-performers are treated similarly in the SP algorithm for the true problems. The average total contribution of a low-performer was 0.255 while the average total contribution of a high-performer was 0.259. There was no significant difference in high-performers' and low-performers' average total contributions on the 17 true problems in the dataset, $t(16) = 0.506$, $p = 0.62$. On the false problems (the top right panel), the average total contribution of a low-performer was 0.281, while the average total contribution of a high-performer was 0.383. High-performers therefore had significantly higher average total contributions than low-performers on the 33 false problems in the dataset, $t(32) = 9.26$, $p < .001$.

As seen in the bottom set of panels of Figure 4, high-performers have a higher average total contribution in the SC algorithm than low-performers for both true and false problems. On the true problems, the average total contribution of a low-performer was 0.116 while the average total contribution of a high-performer was 0.154. High-performers had a significantly higher total contribution than low-performers on the true problems, $t(16) = 5.35$, $p < .001$. On the false problems, the average total contribution of a low-performer was 0.141 while the average total contribution of a

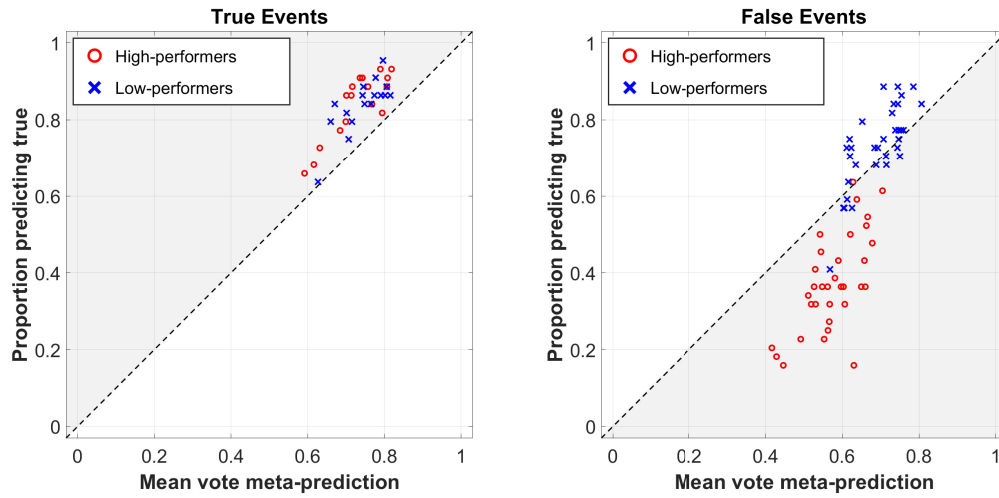
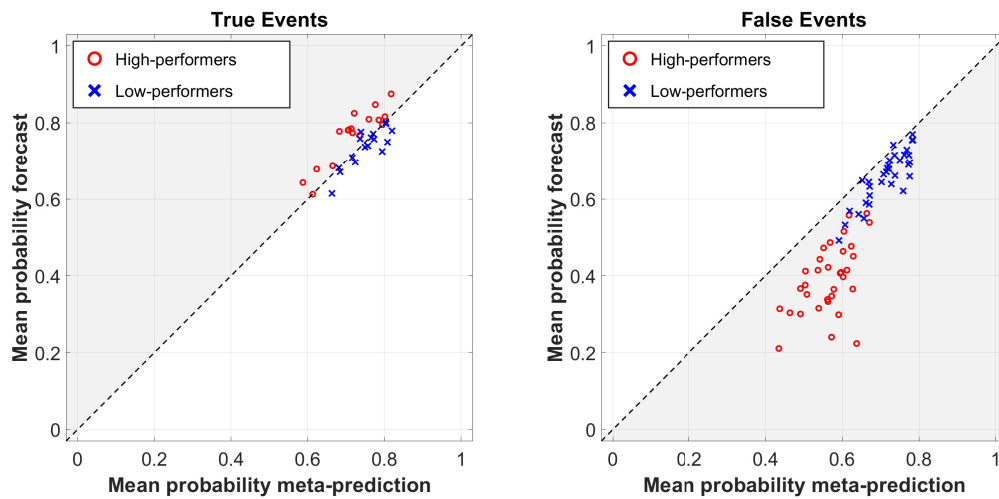
SP Inputs: States Data**SC Inputs: States Data**

Figure 4 The mean responses from high-performers (red circles) and low-performers (blue crosses) on each question in the States dataset. The top two panels show each group's mean votes compared to their mean vote meta-predictions on the true problems (left) and false problems (right). The bottom two panels show each group's mean probability forecast compared to their mean probability meta-prediction for the true problems (left) and the false problems (right). The diagonal line indicates where each group's vote (or forecast) is identical to their vote (or probability) meta-prediction. The shaded regions indicate where each algorithm would generate correct predictions.

high-performer was 0.247. High-performers therefore also had significantly higher total contribution than low-performers on the false problems, $t(32) = 9.85$, $p < .001$.

In Appendix D, we explore an alternative specification where we divide forecasters into quartiles. Consistent with the results here, forecasters in the best-performing quartile have a higher weight in the SC algorithm than the SP algorithm while forecasters in the worst-performing quartile have a lower weight in the SC algorithm than the SP algorithm.

Taken together, the data from the first experiment supports the results from the theoretical model. The weights in the SP algorithm are decreasing as a participant's probabilistic forecast moves away from the uninformed posterior of 0.5 and the algorithm corrects for bias by generating a discontinuity in the weight function at 0.5. This gap ensures that the total contribution of high-performers exceeds that for low-performers on the false problems, but there is no statistically significant difference for true problems.¹⁵ By contrast, the SC algorithm has weights that are increasing as the probabilistic forecast moves away from the estimated prior. As a result, high-performers are over-weighted by the SC algorithm in both true and false problems.

3.2. Experiment 2

Our theoretical model suggests that the performance of the SP algorithm may vary with the proportion of experts and non-experts in the crowd. To create variation in these proportions, our second experiment explores how the relative performance of the SP algorithm and SC algorithm changes with task difficulty.

3.2.1. Methods. We generated 500 science statements at a US primary and secondary grade school level. Problems were adapted from worksheets on the Education Quizzes website (<http://www.educationquizzes.com/us>), and then converted into true or false statements. Approximately 2-3 problems were taken from each worksheet from the Biology, Chemistry, Geography, Physics, and General Science categories, spanning from grades 1 to 12, broken up into five levels of difficulty (grades 1 and 2; grades 3, 4, and 5; grades 6, 7, and 8; grades 9 and 10; and grades 11 and 12). We coded "difficulty 1" as the easiest difficulty, and "difficulty 5" as the hardest difficulty. We treated each set of 100 problems of the same difficulty as an individual dataset.

We recruited 500 respondents from Amazon Mechanical Turk; only respondents inside the US were able to participate in the experiment. Participants were paid a flat fee of \$4.00 for completing the survey. The survey was conducted on the Qualtrics platform. Participants were asked to answer each question as honestly as they could, and were asked not to cheat (e.g., by looking up any of the

¹⁵ This result makes sense in light of the weights shown in Figure 3. As seen there, the prior is 0.74 and a forecaster with no information will have a weight that is smaller than a forecaster who knows with certainty that the answer is false but larger than a forecaster who knows with certainty that the answer is true.

problems online). There were 41 individuals who had reported cheating at the task or had failed to complete the survey. These people were excluded from the analyses and analyses were conducted on the data of the remaining 459 participants.

Participants completed 100 trials each, with each trial comprising one statement which was either true or false, and then followed by the same questions we asked in Experiment 1. Half the statements at each level of difficulty were true, and the other half were false. Each participant saw 20 statements from each level of difficulty, and statements were presented in one of five randomised orders. Participants who took part in any of our previous experiments were excluded from participating. Data collection for all five datasets was completed in July 2019.

Unlike Experiment 1, we did not force participants' probabilistic forecast to match their votes. Instead, participants who provided votes that were inconsistent with their probability forecasts (i.e., voting "true" but predicting a probability of less than 50% of the statement being true, or voting "false" but predicting a probability of greater than 50% of the statement being true) were excluded from the analysis for that particular question. Approximately 11.3% of responses in the dataset were excluded for this reason.¹⁶

In Appendix D, we show that the shape of the weight functions and the relative weighting of the algorithms is similar to the results in Experiment 1. Here, we concentrate on how the two algorithms treat experts. We again ranked and sorted forecasters based on their mean accuracy computed using leave-one-out cross validation. We performed a median split between the best-performing individuals and the worst-performing individuals. This exercise was performed for each grades dataset separately and for all five datasets combined. In the analysis for individual grades, mean accuracy was computed using data only from the particular grade. We then computed and plotted the average contribution of high-performers and low-performers for the SP and SC algorithm on the test problems. We find the following:

Result 3 *In the quiz data, the average total contribution of a high-performer is statistically significantly greater than that of a low-performer in both the SP and SC algorithms for both true and false problems.*

Figure 5 shows the average total contribution of high-performers for each algorithm on each dataset. Aggregating across all 500 problems in the dataset, high-performers had a larger average total contribution than low-performers in both the SP algorithm and SC algorithm. For the SP algorithm, low-performers had an average total contribution of 0.228 whereas high-performers had

¹⁶ We planned to remove inconsistent forecasters prior to running the experiment. However, as the proportion of omitted decisions is relatively large, we also checked to see how both algorithms behave in the full sample. The only substantive difference is noted in footnote 18 below.

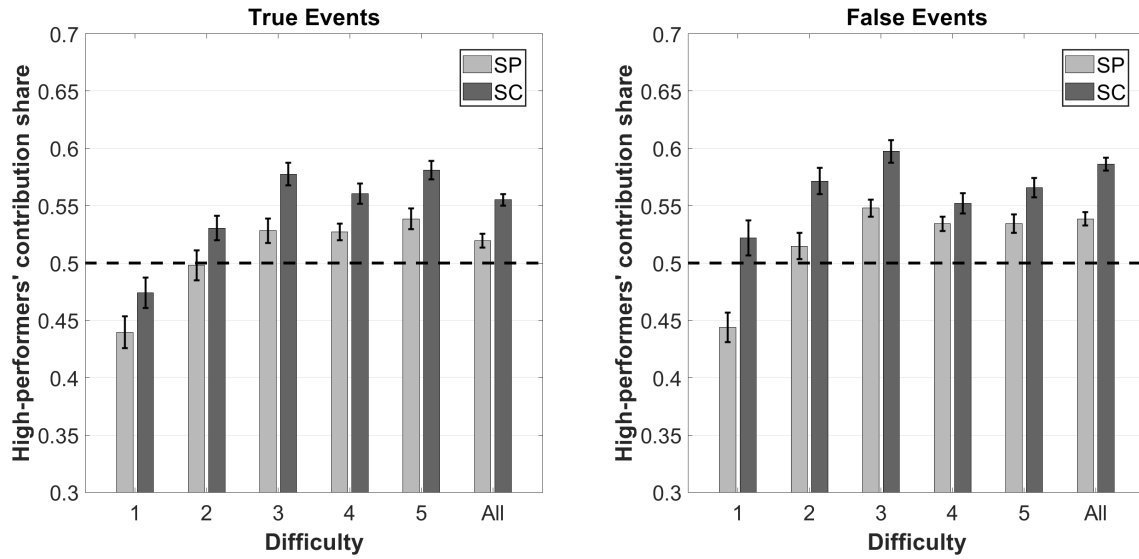


Figure 5 High-performers' average contributions to the SP algorithm and SC algorithm for each of the five individual difficulties and overall across all five difficulties in the quiz dataset. The left panel shows high-performers' share of the crowd contribution on the true events and the right panel shows high-performers' share of the crowd contribution on the false events. The dotted line indicates where high-performers and low-performers have equal contributions to each algorithm's decision.

an average total contribution of 0.272; the difference between average total contributions of high-performers and low-performers is significant, $t(499) = 11.6$, $p < .001$. For the SC algorithm, low-performers had an average total contribution to the SC algorithm of 0.098 whereas high-performers had an average total contribution of 0.134. The difference between average total contributions of high-performers and low-performers is also significant, $t(499) = 18.4$, $p < .001$.

At the dataset level, high-performers had higher contributions than low-performers in the SP algorithm on all but the easiest difficulty. The SC also assigned greater weights to high-performers than low-performers in all but the easiest difficulty. However, we can see that at both the dataset level and the aggregate level, high-performers' contributions to the SC algorithm (relative to low-performers' contributions) were larger than their contributions to the SP algorithm.

3.3. Performance of the SP and SC algorithms

Thus far we have seen that the SC algorithm leverages informed forecasters both theoretically and empirically and assigns larger weights to forecasters who are correct most often and smaller weights to forecasters who are correct least often. In this section, we study whether these properties translate into improved prediction performance.

To assess prediction performance, we use Matthews correlation coefficient (MCC) as our assessment criterion. This criterion takes into account the large number of false problems in the states

dataset, but is similar to accuracy in the quiz datasets where the number of true and false problems are equal. In addition to the SP algorithm and majority voting, we also report the performance of two alternative algorithms that use confidences: the traditional “confidence-weighted” algorithm, which calculates the average probability forecast and assigns a prediction of true if this forecast exceeds 0.5 and a prediction of zero otherwise, and the max-confidence algorithm, which calculates the average half-range confidence of forecasters who predict true and the average half-range confidence of forecasters who predict false and predicts the larger of these two values.¹⁷

We use 95% Confidence Intervals (CIs) to test whether there was a statistically significant difference in performance between the SC and the other algorithms. We compute 95% CIs for the mean difference in MCC between the SP algorithm and the SC algorithm for each of the six datasets from Experiment 1 and 2 and in the aggregate over the five quiz datasets. We find the following result:

Result 4 *The SC algorithm has similar performance to the SP algorithm in the States dataset and outperforms the SP algorithm in the Quiz dataset. The performance of the SC algorithm in the Quiz dataset is driven by high accuracy in the more difficult decision problems. By contrast, the SC algorithm performs poorly relative to other algorithms in the easiest decision problems.*

Support for Result 4 is given in Figure 6, which shows the MCC for the SP and SC algorithm relative to the other three algorithms tested. The SC algorithm significantly outperformed the SP algorithm on Grades 3 dataset (95% CI: [.019, .320]) and Grades 4 dataset (95% CI: [.105, .340]). There was no significant difference in performance between the SC algorithm and SP algorithm on States dataset (95% CI: [-.210, .224]), the Grades 2 dataset (95% CI: [-.078, .213]), or the Grades 5 dataset (95% CI: [-.017, .340]). The SP algorithm significantly outperformed the SC algorithm on the Grades 1 dataset (95% CI: [.331, .050]).

In Figure 4, the rightmost set of bars shows the performance of each algorithm after aggregating across all five quiz datasets from Experiment 2. The SC algorithm appears to outperform the SP algorithm, majority vote, and confidence-weighted algorithms by approximately 0.1 in MCC. Computing the paired mean difference in MCC between the SC algorithm and each other algorithm

¹⁷ Although we concentrate on the SP and SC algorithm, it is useful to briefly describe the properties of these alternative mechanisms. Majority voting assigns an equal weight to all forecasters and is guaranteed to leverage experts only in symmetric problems. The confidence-weighted algorithm assigns larger weights to forecasters with more informative private signals in unbiased problems but not biased ones. It is only guaranteed to predict the correct answer in large samples in unbiased forecasting problems and is sensitive to specific overconfidence. Max confidence has the property that the prediction can switch from true to false when the confidence report of one of the forecasters increases. Thus, it isn't possible to represent the algorithm as a weighted average of reports above and below a single posterior threshold. The algorithm is also not guaranteed to leverage experts in any class of decision problems and is not guaranteed to correctly predict the correct answer in large samples.

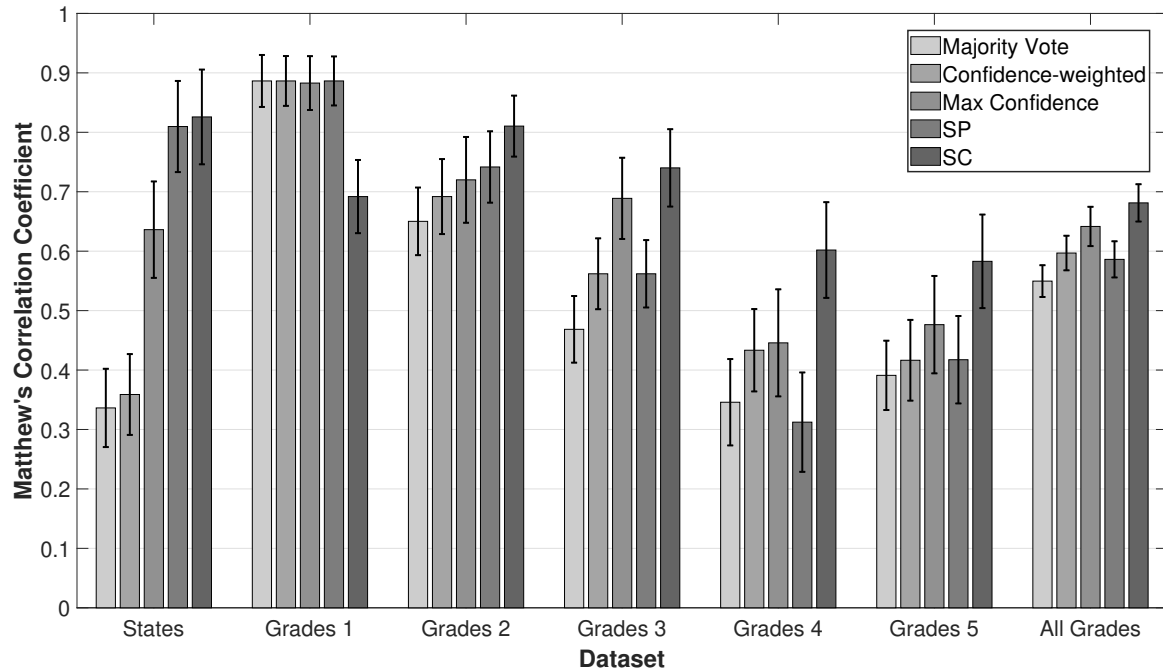


Figure 6 Classification performance of algorithms measured by percentage accuracy on each dataset from Experiment 1 and 2. Error bars show standard error.

over the five quiz datasets, we find that the SC algorithm had a significantly higher MCC than the SP algorithm (95% CI: [.022, .166]), majority voting (95% CI: [.056, .204]) and the confidence-weighted algorithm (95% CI: [.011, .156]).¹⁸ The SC algorithm also has a higher MCC than the max confidence algorithm across all five grade levels, but the difference is not significant (95% CI: [-.039, .120]).

The superior performance on the SC algorithm in hard problems makes sense in the context of the weight functions. In hard problems, many forecasters will have posteriors that are close to the uninformed posterior of 0.5. These forecasters will have large weights in the SP algorithm and this is likely to crowd out the signal from the small number of experts who are likely to exist when the problem is difficult. By contrast, in the SC algorithm, uninformed forecasters will have a zero weight when the prior is unbiased while the small number of expert forecasters in the crowd are likely to have large weights due to having better information.

The relatively poor performance of the SC algorithm in the Grades 1 dataset suggests that the algorithm may not perform well in very easy problems where almost all participants correctly predict the correct answer. We believe this is due to the way that the algorithm handles biased

¹⁸ One caveat to this result is that the exclusion of individuals who were inconsistent appears to negatively impact the performance of the SP algorithm more than the SC algorithm. When inconsistent forecasters are included, the SC algorithm still generally outperforms the SP algorithm and the patterns described above remain. However, the overall difference between the SC and SP algorithm is not statistically significant (95% CI: [-.066, .075]).

priors. By construction, the SC algorithm eliminates common knowledge from the algorithm by setting the weight of an individual who receives no private signal to zero. If the common knowledge is indeed informative and there is little additional private information, the algorithm may perform poorly relative to other algorithms. In contrast, the SP algorithm assigns large weights to forecasters who have little private information, and therefore does not eliminate forecasters with common information.

Although it is not the focus of this paper, our theoretical results show that both the SP and SC algorithm are able to correct for bias in large samples and thus any hybrid algorithm that selects between them based on a secondary criterion will also correctly predict the correct answer. Our data suggests that the SC algorithm performs well on hard problems while the SP algorithm does well on problems that are easy. In principle, an algorithm that switches between these two algorithms based on task difficulty may do better than either algorithm alone. For instance, across our states and quiz datasets, an algorithm that uses the SC algorithm's prediction when the average probability meta-prediction is between $[0.3, 0.7]$ and the SP algorithm's prediction in other circumstances has a MCC of .72 in the quiz data, which significantly outperforms both the SC algorithm (95% CI: $[-.003, .077]$) and the SP algorithm (95% CI: $[-.066, .200]$). It also outperforms both the SP and SC algorithm in the states data, but the difference is not statistically significant (95% CI: $[-.057, .257]$ and $[-.070, .259]$, respectively).

4. Conclusion

Modern forecasting algorithms use the Wisdom of Crowds to produce forecasts better than those of the best identifiable expert. However, these algorithms may be inaccurate when crowds are systematically biased or when expertise varies substantially across forecasters. Recent work by Prelec et al. (2017) has shown that meta-predictions—a forecast of the average forecast of others—can be used to correct for biases even when no external information such as a forecasters past performance is available. Our paper explored how meta-predictions can also be used to improve predictions by identifying and leveraging expertise in the crowd.

We began by outlining an alternative confidence-based version of the SP algorithm. This algorithm retains the theoretical property that it will always predict the correct answers in large samples even when forecasters have a biased prior. In contrast to the SP algorithm, we showed that the SC algorithm weights individuals with more informative private signals more than those with less informative private signals. The algorithm also leverages expertise and can mitigate biases in confidences that arise when individuals who believe the consensus position is correct are overconfident and individuals who believe the consensus position is incorrect are under-confident. Over two experiments, we find that the new SC algorithm does a better job in weighting better-informed

forecasters than the original algorithm and show that individuals with higher mean accuracy contribute more to the algorithm than other forecasters.

We also explored the properties of the SP and SC algorithm across a range of problems that varied in difficulty. Overall, the SC algorithm was more effective at leveraging expertise than the SP algorithm. However, the efficacy of the weights did not translate into improved performance at all levels of difficulty. On the easiest problems, the SC algorithm was significantly worse than the SP algorithm, despite the SC algorithm leveraging expertise more effectively. In contrast, the SC algorithm was generally more effective than the SP algorithm on the moderate-to-hard problems. Thus, despite the theoretical advantages of the SC algorithm, the empirical performance of these algorithms suggests they may be suited to different types of problems, rather than being strictly better or worse than one another.

Overall, our theoretical and empirical findings provide useful insight into how these algorithms can be used to leverage expertise in the single-question domain. The weights used by the SC algorithm have useful properties relating to forecasters' expertise, but importantly, the properties of these weights are not fundamentally tied to each algorithm. Thus, the weights of the SC algorithm can be used independently, for example, for the purposes of improving forecasts in the probabilistic domain (Martinie et al. 2019), or for other purposes such as identifying high-performing individuals for the purposes of compensation or evaluation.

There exist other algorithms that seek to identify expertise in the single question domain, such as those based on forecasters' confidence (Koriat 2008) or decision similarity (Kurvers et al. 2019). These other measures are most effective in 'kind' environments or low-difficulty problems, where the majority of forecasters are likely to vote correctly (Koriat 2008, Kurvers et al. 2019). In contrast, our results suggest that the SC weights are better suited for identifying expertise on moderate-to-high difficulty problems, where the majority of forecasters may often be biased and vote incorrectly. Our results are therefore complementary to the existing literature in that they can be used to identify and leverage expertise in different forecasting environments.

References

- Baillon A, Tereick B, Wang TV (2020) Follow the money, not the majority: Incentivizing and aggregating expert opinions with Bayesian markets, mimeo.
- Blackwell D (1951) Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (The Regents of the University of California).
- Blackwell D (1953) Equivalent comparisons of experiments. *The Annals of Mathematical Statistics* 265–272.
- Blackwell D, Girshick MA (1979) *Theory of games and statistical decisions* (Courier Corporation).

- Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Science* 61(2):267–280.
- Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5(4):559–583.
- Condorcet Md Marie Jean Antoine Nicolas de Caritat (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (De l'Imprimerie royale).
- Cooke RM (1991) *Experts in uncertainty: opinion and subjective probability in science* (Oxford University Press on Demand).
- Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, Johannesson M (2015) Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* 112(50):15343–15347.
- Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological review* 101(3):519.
- Fischhoff B, MacGregor D (1982) Subjective confidence in forecasts. *Journal of Forecasting* 1(2):155–172.
- Galton F (1907) Vox populi (the wisdom of crowds). *Nature* 75(7):450–451.
- Genre V, Kenny G, Meyler A, Timmermann A (2013) Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting* 29(1):108–121.
- Gigerenzer G (1984) External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology* 97(2):185–195.
- Gigerenzer G, Hoffrage U, Kleinbölting (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review* 98(4):506–528.
- Gillen B, McKenzie J, Plott CR (2018) Two information aggregation mechanisms for predicting the opening weekend box office revenues of films: Boxoffice prophecy and guess of guesses. *Economic Theory* 65(1):25–54.
- Griffin D, Brenner L (2007) Perspectives on probability judgment calibration. Koehler DJ, ed., *Blackwell Handbook of Judgment and Decision*, 177–199 (John Wiley & Sons, Incorporated).
- Hertwig R (2012) Tapping into the wisdom of the crowd—with confidence. *Science* 336(6079):303–304.
- Hintzman DL, Nozawa G, Irmscher M (1982) Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning and Verbal Behavior* 21(2):128–141.
- Koriat A (2008) Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(4):945.
- Koriat A (2012) When are two heads better than one and why? *Science* 336(6079):360–362.
- Kurvers RH, Herzog SM, Hertwig R, Krause J, Moussaid M, Argenziano G, Zalaudek I, Carney P, Wolf M (2019) How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances* 5(11):eaaw9011.

- Lee MD, Lee MN (2017) The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making* 12(4):328.
- Lieberman V, Tversky A (1993) On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin* 114(1):162–173.
- Marschak J, Miyasawa K (1968) Economic comparability of information systems. *International Economic Review* 9(2):137–174.
- Marschak J, Radner R (1972) *Economic Theory of Teams (Cowles Foundation Monograph 22)* (Yale University Press, New Haven, CT).
- Martinie M, Wilkening T, Howe P (2019) Using meta-predictions to identify experts in the crowd when past performance is unknown. *PLoS ONE* 15(4):1–11.
- McCoy J, Prelec D (2017) A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint arXiv:1703.04778* .
- Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, Baker J, Hou Y, Horowitz M, Ungar L, Tetlock P (2015) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3):267–281.
- Müller-Trede J, Choshen-Hillel S, Barneron M, Yaniv I (2017) The wisdom of crowds in matters of taste. *Management Science* 64(4):1779–1803.
- Palley A, Satopää V (2020) Boosting the wisdom of crowds within a single judgment, problem: selective averaging based on peer predictions, available at <https://ssrn.com/abstract=3504286>.
- Palley A, Soll JB (2018) Extracting the wisdom of crowds when information is shared. *Management Science* 65(5):1949–2443.
- Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532.
- Satopää VA, Pemantle R, Ungar LH (2016) Modeling probability forecasts via information diversity. *Journal of the American Statistical Association* 111(516):1623–1633.
- Simmons JP, Nelson LD, 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.
- Surowiecki J (2005) *The wisdom of crowds* (Anchor).
- Tereick B (2019) Improving information aggregation through meta-cognition judgments, mimeo.
- Tetlock PE (2017) *Expert political judgment: How good is it? How can we know?* (Princeton University Press).

Online Appendix A: The relationship between weights and accuracy

In the main text, we provided Proposition 3 that showed that the expected weight of the SP and SC algorithms were ordered. This implies that the means of the limiting distributions are well ordered. We also noted that the variances of the two algorithms are not ordered. To have a better sense of how the two algorithms are likely to perform in small and moderate samples, we construct 100,000 randomly generated unbiased information systems using the following process. First, in each state, we draw five uniform $[0, 1]$ variables, $x_1, \dots, x_5$, and set $Q_{oi} = \frac{x_i}{\sum_{i=1}^5 x_i}$. By construction, the elements of each row sum to one and there will be both a state where $Q_{Ti} > Q_{Fi}$ and a state where $Q_{Ti} < Q_{Fi}$. Thus, each information system will be responsive.

We next generate 1000 samples of size 100 to calculate the mean ($\bar{W}$) and variance ($Var(W)$) of the sample in both the case where the correct answer is true and the case where the sample is false. We use samples of 100 to ensure that the variance generated in re-weighting the observations in each algorithm is taken into account. We also chose this sample size because it is the sample used in our experiments.

Over the 100,000 samples, the average value of $\bar{W}^{SC}$ is 0.741 and the average value of $\bar{W}^{SP}$ is 0.674 when the state is true. The average observation-level coefficient of variance, $\frac{100Var(W)}{\bar{W}}$, of the SC algorithm is 0.366 while the average observation-level coefficient variance of the SP algorithm is 0.378. Both the difference in means and the difference in the coefficient of variance are significantly different from zero, though the magnitude difference in the coefficient of variances is very small (paired t-test of means: $t(99999) = 376.0$, $p < .001$; paired t-test of coefficient of variance: $t(99999) = 32.5$, $p < .001$). The results for false question are nearly identical ($1 - \bar{W}^{SC} = 0.740$ and $1 - \bar{W}^{SP} = 0.674$ when the state is false; coefficient of variances (using a mean of $1 - \bar{W}$ as the denominator) are 0.366 and 0.378 respectively).

Based on the mean and the variance generated by the sample, we approximate the sample size N necessary to ensure an accuracy of 97.5% by finding the point where the lower bound of the confidence interval is equal to 0.5 for both the case where the answer is true and false:

$$\frac{1}{2} = \bar{W} - 1.96 \frac{(100Var(W))^{\frac{1}{2}}}{N^{\frac{1}{2}}}.$$

The maximum of these two N is the estimated sample size necessary to generate an accuracy of 97.5% for both algorithms.

Across the 100,000 samples, the SC algorithm is predicted to require a smaller N in 99.1% of cases while the SP algorithm is predicted to have a smaller sample in 0.62% of cases. Restricting attention to the 76,731 cases where at least one algorithm requires a sample size of at least 30 and where the central limit theorem is likely to be a reasonable approximation, the SC algorithm is predicted to require a smaller sample size in 99.9% of cases. These results suggest that the SC algorithm is likely to be more efficient in the vast majority of unbiased decision problems in cases where all forecasters are Bayesian.

Figure 7 plots the minimum number of individuals necessary to ensure that the SP and SC algorithms generate the correct forecast 97.5% of the time for two information systems containing experts who know the correct state and uninformed novices. In both panels, forecasters are drawn from a population where a proportion θ are experts and are fully informed about the correct answer and $1 - \theta$ are novices. In the

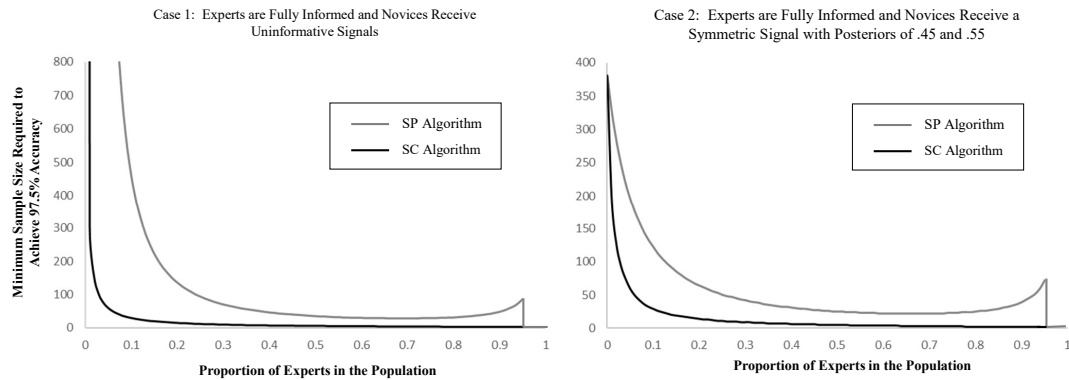


Figure 7 The left panel shows the sample size necessary to achieve 97.5% accuracy with the SP and SC algorithm from a population consisting of fully informed experts and uninformed novices. The right panel shows the sample sizes necessary to achieve 97.5% accuracy with the SP and SC algorithms when novices receive a symmetric signal that generate posteriors of 0.55 and 0.45.

left hand panel, novices have no informative signal and both algorithms will be correct 50% of the time when $\theta = 0$ for any N . In the right hand panel, each novice receives an independent signal that is correct 55% of the time and incorrect 45% of the time. In both graphs, we concentrate on a symmetric information system where $s_\emptyset = .5$. The cutoffs reported are derived analytically using the exact sample distribution or a normal approximation in cases where the Lindeberg-Lévy Central Limit Theorem applies. We randomly pick a predicted state in cases where either algorithm returns an indeterminate value.

As seen in the left hand panel, the SC algorithm requires a very small sample sizes to accurately predict the correct answer when novices are fully uninformed. This is because uniformed forecasters have zero weight in the algorithm and it only takes a single informed forecaster to generate the correct answer. In the SP algorithm, by contrast, uninformed individuals have a larger weight than the informed forecasters for any θ . Although the expected contribution of each novice is zero, they nonetheless create substantial noise in the algorithm that can lead to inaccurate predictions. As a result, the SP algorithm requires a larger sample than the SC algorithm for any proportion of experts. The difference in required sample sizes is particularly pronounced for cases where the proportion of experts is small. For example, when only 10% of the population is an expert, the SC algorithm requires a sample of 29 participants to ensure an accuracy of 97.5% while the SP algorithm requires a sample of 462.

The right hand panel shows that the SP algorithm continues to require larger sample sizes even when the novices are partially informed and that the SC algorithm requires a smaller sample for any proportion of experts. This graph shows that the difference in sample sizes seen in the left hand panel is not due to the assumption that novice forecasters were fully uninformed.

Online Appendix B: Expertise and the SP algorithm

In this appendix, we explore how the SP algorithm treats experts. In part 1 we provide counter examples that show that the SP algorithm does not always leverage experts in a variety of information systems. We then

provide two additional conditions on the information system that are sufficient to ensure that the algorithm leverages experts in symmetric information systems. Finally, we provide an example that highlights some of the intuition that underlines the proof and discuss the cases where we expect the SP algorithm to perform best when forecasters are heterogeneous in their expertise.

Part 1: Counter Examples

Examples 1 and 2 below show that in cases where the prior or the posteriors are asymmetric, it is possible to find counter examples where the expected total contribution of experts in the SP algorithm is less than that of novices in at least one state.

EXAMPLE 1. Consider an environment where $\theta = .5$, the prior $s_{\emptyset} = .8$ and where the set of additional posteriors are $\{0, .4, .6, 1\}$. Suppose further that the experts' information service over $\{0, .4, .6, .8, 1\}$ is

$$Q^E = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and the Novices' information service is

$$Q^N = \begin{bmatrix} 0 & 0 & .375 & 0 & .625 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Note that the Novices will always vote for True regardless of the state because their posteriors are always greater or equal to .6. Thus, this is a group that is biased and information will only influence their meta-predictions.

We now show that the expected total contribution of expert is *not* greater than the expected total contribution of the novices in the true state. For the experts,

$$\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E, T) = 1 - (.5 * 1 + .5 * 1) = 0,$$

while for the Novices

$$\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T) = 1 - (.5 * .85 + .5 * 1) = .075.$$

Example 1 shows that when the prior is biased, the expected contribution of an expert may be smaller than that of the novice for at least one of the states. The following example shows that even when the prior is unbiased, it is still possible to construct information systems where the expected total contribution of an expert is less than that of a novice.

EXAMPLE 2. Consider an environment where $\theta = .5$, the prior $s_{\emptyset} = .5$ and where the set of additional signal realizations are $\{\frac{x}{x+1}, 1\}$ with $x \in [0, 1)$. Suppose further that the experts' information service over $\{\frac{x}{x+1}, .5, 1\}$ is

$$Q^E = \begin{bmatrix} x & 0 & 1-x \\ 1 & 0 & 0 \end{bmatrix}$$

and the Novices' information service over $\{\frac{x}{x+1}, .5, 1\}$ is

$$Q^N = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

We now show that the expected total contribution of an expert may be lower than the expected total contribution of a novice and that it can be negative for x close to 1. For the experts,

$$\mathbb{E}V(Q^E|T) = 1 - x$$

and

$$\mathbb{E}M^V(\theta|Q^E, T) = .5M^V(Q^E|Q^E, T) + .5M^V(Q^N|Q^E, T) = \frac{1-x}{2(1+x)} + .25$$

In the limit, as $x \rightarrow 1$

$$\lim_{x \rightarrow 1} [\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E, T)] = 0 - .5^2 = -.25.$$

This is strictly below $\lim_{x \rightarrow 1} [\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T)] = .25$.

Part 2: Sufficient Conditions

We now discuss two additional properties of the information systems that are sufficient to ensure that the SP algorithm leverages expertise in symmetric information systems.

Strict Garbling: When information system Q^E is more informative than Q^N , we can find a garbling matrix Z such that each signal in Q^N can be found by adding noise to the signals in Q^E . To guarantee that the SP algorithm leverages expertise, we will require a stronger condition. Rather than using any set of signals from Q^E , we will require that Q^N can be constructed by garbling only signals in Q^E that are at least as informative as the signal being constructed in Q^N . Let $\hat{s} \in [0, 0.5]$ be an arbitrary posterior between 0 and 0.5. Further, let

$$F^t(\hat{s}) := \sum_{\{i|s_i \leq \hat{s}\}} [Q_{Ti}^t + Q_{T(m+2-i)}^t]$$

to be the probability of having a posterior that is less than or equal to an arbitrary posterior $\hat{s}$ or greater than or equal to $1 - \hat{s}$ when receiving signals from information system $t \in \{E, N\}$ in state T . Note that in a symmetric information system,

$$\sum_{\{i|s_i \leq \hat{s}\}} [Q_{Ti}^t + Q_{T(m+2-i)}^t] = \sum_{\{i|s_i \leq \hat{s}\}} [Q_{Fi}^t + Q_{F(m+2-i)}^t]$$

and thus, under symmetry, $F^t(\hat{s})$ is invariant to the state chosen to evaluate it.

DEFINITION 7. Q^N is a **strict garbling** of Q^E if (i) both Q^N and Q^E are symmetric, (ii) $F^N(\hat{s}) \leq F^E(\hat{s})$ for all $\hat{s} < 0.5$ and (iii) exists at least one $\hat{s}$ for which $F^N(\hat{s}) < F^E(\hat{s})$.

Problem Difficulty: Forecasting problems are the most difficult when forecasters receive weak signals about the true state and where the vote shares are close to 50:50. We define a forecasting problem as hard if at least a quarter of the population will answer the question incorrectly:

DEFINITION 8. A forecasting problem is **hard** if less than 75% of forecasters vote “true” in the true state and greater than 25% of forecasters vote “true” in the false state.

The following result provides a set of sufficient conditions that ensure that the SP algorithm leverages expertise in environments where Assumptions 1-3 hold and where there are exactly two information systems:¹⁹

PROPOSITION 6. *The SP algorithm leverages expertise if information systems Q^N and Q^E are symmetric, Q^N is a strict garbling of Q^E , and the forecasting problem is hard.*

Although the proof of proposition 6 is technical, it is again related to the slope and level of the vote meta-prediction function. We demonstrate this here with a simple example.

Consider a decision problem where the prior $s_\varnothing = .5$ and where the set of additional posteriors are $\{0, .4, .6, 1\}$. Suppose further that the experts' information service over $\{0, .4, .5, .6, 1\}$ is

$$Q^E = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and the Novices' information service over $\{0, .4, .5, .6, 1\}$ is

$$Q^N = \begin{bmatrix} 0 & .4 & 0 & .6 & 0 \\ 0 & .6 & 0 & .4 & 0 \end{bmatrix}$$

In this problem, experts know the correct state, while novices have weak but correctly informative signals. We study how the expected total contributions of novices and experts change with θ .

Since Q^N and Q^E are symmetric, we will restrict attention to the true state. First, note that

$$\begin{aligned} M^V(\theta|s_k) &= \theta M^V(Q^E|s_k) + (1 - \theta)M^V(Q^N|s_k) \\ &= \theta s_k + (1 - \theta)[.6s_k + .4(1 - s_k)]. \end{aligned}$$

For an expert, $\mathbb{E}[V(Q^E|T)] = 1$, and $\mathbb{E}[M^V(\theta|Q^E, T)] = M^V(\theta|s_k = 1) * 1 = \theta + .6(1 - \theta)$. Thus

$$\mathbb{E}[V(Q^E|T)] - \mathbb{E}[M^V(\theta|Q^E, T)] = .4 - .4\theta. \quad (3)$$

For the novice, $\mathbb{E}[V(Q^N|T)] = .6$, $M^V(\theta|s_k = .4) = .4\theta + (1 - \theta)[.6 * .4 + .4 * .6]$, and $M^V(\theta|s_k = .6) = .6\theta + (1 - \theta)[.6^2 + .4^2]$. Thus

$$\begin{aligned} \mathbb{E}[M^V(\theta|Q^N, T)] &= .4M^V(\theta|s_k = .4) + .6M^V(\theta|s_k = .6) \\ &= .4[.4\theta + .48(1 - \theta)] + .6[.6\theta + .52(1 - \theta)] \\ &= .52\theta + .504(1 - \theta). \end{aligned}$$

It follows that

$$\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T) = .6 - [.52\theta + .504(1 - \theta)] = .096 - .016\theta. \quad (4)$$

Using equations (3) and (4) above, the total contributions of an expert exceeds that of a novice when $.4 - .4\theta > .096 - .016\theta$ or, equivalently, when $\theta < \frac{19}{24}$.

¹⁹ We note that unlike the proof for the SC algorithm, Proposition 6 does not necessarily generalize to cases where there are more than two information systems.

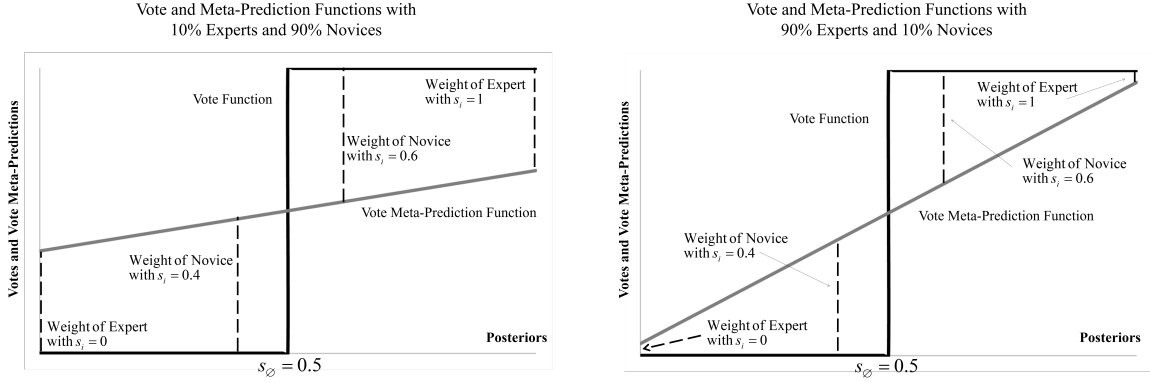


Figure 8 The left panel shows the vote function and vote meta-prediction over all possible posteriors for the case where 10% of the population are experts and 90% are novices. The right panel shows the vote function and vote meta-prediction over all posteriors for the case where 90% of the population are experts and 10% are novices. Weights are equal to the absolute distance between the two functions.

Figure 8 plots the meta-prediction line $M^V(\theta|s_k)$ and the vote function for $\theta = 0.1$ and $\theta = 0.9$. As seen on the left hand side, when $\theta = 0.1$, the meta-prediction line is relatively flat and the weight given to the expert is similar to that of the novices. Because the votes of the novices tend to cancel out while all experts perfectly predict the correct state, experts are leveraged in the decision problem.

By contrast, when $\theta = 0.9$, the slope of the meta-prediction line is close to one and the weights of individuals with high-quality signals grow small. Thus, although some of the forecasts of the novices partially cancel out, the expected contribution of the experts falls below that of the novices.²⁰

Within the class of symmetric problems, the SP algorithm is likely to perform best when the total weight given to experts in the algorithm is largest. Our simple example shows that when there are too many experts who know the correct state, the weights given to each individual expert may grow small. Thus, our analysis suggests that the SP algorithm is likely to do best in cases where there are an intermediate number of experts. This is the case with the example above, where the difference between the total expected contribution of all experts, $\theta[\mathbb{E}[V(Q^E|T)] - \mathbb{E}[M^V(\theta|Q^E, T)]]$, and the total expected contribution of all novices, $(1 - \theta)[\mathbb{E}[V(Q^N|T)] - \mathbb{E}[M^V(\theta|Q^N, T)]]$, is largest at $\theta \approx .616$.

Online Appendix C: Proofs

In this appendix we provide proofs for all the lemmas and propositions in the paper. We provide the proofs to Lemmas (1) – (4) first before presenting the proofs for Propositions (1) – (6).

Lemmas (1) – (4)

Proof of Lemma 1: In this proof, we show that the Surprisingly Popular (SP) algorithm of PSM can be rearranged such that each forecaster's vote is weighted by the normalized, absolute difference between their vote and meta-prediction. We begin with the original form of the SP algorithm and rearrange it to

²⁰ Note that when $\theta = .9$, 96% of forecasters will vote for the right answer and the problem is not classified as hard. Thus, our sufficient conditions do not cover this case.

show that it is identical to a weighted form, where the weights are given by the absolute difference between their vote and their meta-prediction, normalized by the sum of this difference over all forecasters:

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N \frac{|V_i - M_i^V| V_i}{\sum_{j=1}^N |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

In the original SP algorithm, the proportion of the crowd voting for that outcome is compared to the mean meta-prediction, and the most under-predicted outcome is then predicted to be correct. Formally,

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N (V_i - M_i^V) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The crowd for an event with N forecasters can be decomposed into T forecasters who vote true and F forecasters who vote false, $N = T + F$. The report of each forecaster who votes true $t \in \{0, \dots, T\}$, is given by $X_t := (V_t, P_t, M_t^V, M_t^P)$, and the report of each forecaster who votes false, $f \in \{0, \dots, F\}$, is given by $X_f := (V_f, P_f, M_f^V, M_f^P)$. The SP equation can therefore be decomposed into

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T (V_t - M_t^V) + \sum_{f=1}^F (V_f - M_f^V) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Rearranging this, we get

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T (V_t - M_t^V) > - \sum_{f=1}^F (V_f - M_f^V) \\ 0 & \text{otherwise.} \end{cases}$$

As $V_f = 0$, $V_t = 1$, and $M_i^V \in [0, 1]$ the difference between votes and vote meta-predictions for any individual who votes false will always be equal to or less than 0,

$$V_f - M_f^V \leq 0,$$

and the difference between votes and vote meta-predictions for any individual who votes true will always equal or exceed 0,

$$V_t - M_t^V \geq 0.$$

The SP equation is therefore equivalent to

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T |V_t - M_t^V| > \sum_{f=1}^F |V_f - M_f^V| \\ 0 & \text{otherwise.} \end{cases}$$

Adding the terms on the left to both sides, we obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T 2|V_t - M_t^V| > \sum_{t=1}^T |V_t - M_t^V| + \sum_{f=1}^F |V_f - M_f^V| \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$\sum_{t=1}^T |V_t - M_t^V| + \sum_{f=1}^F |V_f - M_f^V| = \sum_{j=1}^N |V_j - M_j^V|,$$

we can collect the terms on the right:

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T 2|V_t - M_t^V| > \sum_{j=1}^N |V_j - M_j^V| \\ 0 & \text{otherwise.} \end{cases}$$

After dividing both sides by the RHS term and dividing both sides by 2, we obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \frac{\sum_{t=1}^T |V_t - M_t^V|}{\sum_{j=1}^N |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

which is identical to

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T \frac{|V_t - M_t^V|}{\sum_{j=1}^N |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Since $V_t = 1$, we can multiply both sides by V_t and simplify the terms on the right to obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T \frac{|V_t - M_t^V| V_t}{\sum_{j=1}^N |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

As $V_f = 0$,

$$\sum_{f=1}^F \frac{|V_f - M_f^V| V_f}{\sum_{j=1}^N |V_j - M_j^V|} = 0,$$

and we can add this summation term to both sides of the previous equation and simplify the terms on the right to obtain

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{t=1}^T \frac{|V_t - M_t^V| V_t}{\sum_{j=1}^N |V_j - M_j^V|} + \sum_{f=1}^F \frac{|V_f - M_f^V| V_f}{\sum_{j=1}^N |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Collecting the terms on the left, we obtain the weighted version of the SP algorithm, thus proving Lemma 1,

$$T_{SP}(X) = \begin{cases} 1 & \text{if } \sum_{i=1}^N \frac{|V_i - M_i^V| V_i}{\sum_{j=1}^N |V_j - M_j^V|} > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Lemma 2: For MLRP to be satisfied, we need to show that for any set of signals that can be drawn with $s_i > s_j$ and $s_k > s_l$,

$$p(s_i|s_k)p(s_j|s_l) > p(s_j|s_k)p(s_i|s_l) \quad (5)$$

Note that a signal can only be drawn if it occurs with positive probability in its information service. Thus $p(s_i) > 0$, $p(s_j) > 0$, $p(s_k) > 0$, and $p(s_l) > 0$.

Assumption 2 implies that

$$p(s_a|s_b) = \frac{p(s_a, s_b)}{p(s_b)} = \frac{p(s_a|T)p(s_b|T)p(T) + p(s_a|F)p(s_b|F)p(F)}{p(s_b)}.$$

For $a \in \{i, j\}$ and $b \in \{k, l\}$. Rearranging Bayes Rule, it is the case that:

$$\frac{p(T)p(s_b|T)}{p(s_b)} = p(T|s_b) = s_b$$

and thus

$$p(s_a|s_b) = p(s_a|T)s_b + p(s_a|F)(1 - s_b) = Q_{Ta}^t s_b + Q_{Fa}^t (1 - s_b). \quad (6)$$

We first prove that MLRP holds for the case where $s_i > s_j > 0$ and $s_k > s_l > 0$. By the construction of the Q matrix

$$s_a = \frac{Q_{Ta}^t p(T)}{Q_{Ta}^t p(T) + Q_{Fa}^t (1 - p(T))}.$$

Under the assumption that $s_j > 0$ and $s_l > 0$, this can be rewritten as

$$Q_{Fa}^t = \frac{1 - s_a}{s_a} \frac{p(T)}{1 - p(T)} Q_{Ta}^t = \frac{1 - s_a}{s_a} \frac{s_{\emptyset}}{1 - s_{\emptyset}} Q_{Ta}^t$$

for $a \in \{i, j\}$. Substituting this into (6) implies that

$$p(s_a|s_b) = Q_{Ta}^t s_b \left[1 + \frac{1 - s_a}{s_a} \frac{1 - s_b}{s_b} \frac{s_{\emptyset}}{1 - s_{\emptyset}} \right]. \quad (7)$$

Let $r_{ab} = \frac{1 - s_a}{s_a} \frac{1 - s_b}{s_b} \frac{s_{\emptyset}}{1 - s_{\emptyset}}$ for $a \in \{i, j\}$ and $b \in \{k, l\}$. Substituting (7) into (5), MLRP is satisfied if:

$$(1 + r_{ik})(1 + r_{jl}) > (1 + r_{il})(1 + r_{jk})$$

Expanding this equation, MLRP is satisfied if:

$$1 + r_{ik} + r_{jl} + r_{ik}r_{jl} > 1 + r_{il} + r_{jk} + r_{il}r_{jk}$$

Next, noting that $r_{ik}r_{jl} = r_{il}r_{jk}$, MLRP is satisfied if

$$r_{ik} + r_{jl} > r_{il} + r_{jk}.$$

Rearranging this equation, MLRP is satisfied if

$$\left[\frac{1 - s_i}{s_i} - \frac{1 - s_j}{s_j} \right] \left[\frac{1 - s_k}{s_k} - \frac{1 - s_l}{s_l} \right] > 0.$$

By the assumption that $s_i > s_j$ and $s_k > s_l$, both terms on the LHS are negative and thus this equation always holds.

We now check the cases for which (i) $s_j = 0$ but $s_l > 0$, (ii) $s_j > 0$ but $s_l = 0$, and (iii) both $s_j = 0$ and $s_l = 0$. When $s_j = 0$ but $s_l > 0$, $p(s_j|s_b) = Q_{Fj}^t(1 - s_b)$ and MLRP is satisfied if

$$[Q_{Ti}^t s_k(1 + r_{ik})][Q_{Fj}^t(1 - s_l)] > [Q_{Ti}^t s_l(1 + r_{il})][Q_{Fj}^t(1 - s_k)].$$

When $s_k = 1$, the RHS is zero and the LHS is strictly positive. Thus MLRP holds. When $s_k < 1$, the equation is equivalent to

$$\frac{s_k}{1 - s_k} > \frac{s_l}{1 - s_l},$$

which is satisfied due to the assumption that $s_k > s_l$.

When $s_l = 0$ and $s_j > 0$, MLRP holds if

$$\frac{Q_{Ti}^t s_k + Q_{Fi}^t(1 - s_k)}{Q_{Tj}^t s_k + Q_{Fj}^t(1 - s_k)} > \frac{Q_{Fi}^t}{Q_{Fj}^t}. \quad (8)$$

If $s_i = 1$, the RHS is equal to zero and the LHS is strictly positive. Thus MLRP holds. When $s_i < 1$, $s_j > 0$, and $s_l = 0$, $Q_{Ti}^t = \frac{s_i}{1 - s_i} \frac{1 - s_o}{s_o} Q_{Fi}^t$ and, after some algebra, the equation is equivalent to

$$\frac{s_i}{1 - s_i} > \frac{s_j}{1 - s_j},$$

which is always satisfied. Thus MLRP holds in this case.

Finally when $s_l = 0$ and $s_j = 0$, MLRP holds if

$$[Q_{Ti}^t s_k + Q_{Fi}^t(1 - s_k)]Q_{Fj}^t > [Q_{Fj}^t(1 - s_k)]Q_{Fi}^t.$$

Rearranging, MLRP holds if $Q_{Ti}^t s_k Q_{Fj}^t > 0$, which is always true.

Proof of Lemma 3: Assume that the event is true. The share of true votes from information service $t \in \{E, N\}$ (given that the state is true) is given by

$$\mathbb{E}V(Q^t|T) = \sum_{\{i|s_i \geq .5\}} \gamma(Q_{Ti}^t),$$

where $\gamma(Q_{oi}^t) = \frac{1}{2}Q_{oi}^t$ if $s_i = .5$ and $\gamma(Q_{oi}^t) = Q_{oi}^t$ otherwise. The meta-prediction of an individual in group t with signal s_k is

$$M^V(Q^t|s_k) = s_k \mathbb{E}V(Q^t|T) + (1 - s_k) \mathbb{E}V(Q^t|F).$$

The expected meta-prediction of forecasters in information service Q^t made by forecasters with information service Q^τ ($\tau = \{N, E\}$) when the state is o is given by

$$\mathbb{E}M^V(Q^t|Q^\tau, o) = \sum_k M^V(Q^t|s_k)Q_{ok}^\tau.$$

Aggregating up across novices and experts, the expected meta-prediction of votes from information service Q^t given state o is

$$\mathbb{E}M^V(Q^t|o) = \theta \mathbb{E}M^V(Q^t|Q^E, o) + (1 - \theta) \mathbb{E}M^V(Q^t|Q^N, o).$$

In the true state, the meta-prediction will underestimate (or be equal to) the true proportion of votes for the true state if for all $t \in \{E, N\}$,

$$\mathbb{E}V(Q^t|T) \geq \mathbb{E}M^V(Q^t|T). \quad (9)$$

We allow for equality here to account for the cases where (i) all individuals know the state is true or (ii) all individuals are uninformed with a prior of $s_{\emptyset} = .5$. In these special cases the expected votes and expected meta predictions will be equal.

Noting that $\mathbb{E}V(Q^t|T) = \theta\mathbb{E}V(Q^t|T) + (1 - \theta)\mathbb{E}V(Q^t|T)$, equation (9) holds if for $t \in \{E, N\}$ and $\tau \in \{E, N\}$,

$$\mathbb{E}V(Q^t|T) \geq \mathbb{E}M^V(Q^t|Q^\tau, T)$$

Next, noting that (i)

$$\mathbb{E}M^V(Q^t|Q^\tau, T) = \sum_k M^V(Q^t|s_k)Q_{Tk}^\tau,$$

(ii) $M^V(Q^t|s_k) = s_k\mathbb{E}V(Q^t|T) + (1 - s_k)\mathbb{E}V(Q^t|F)$, and (iii) $\sum_k Q_{Tk}^\tau = 1$, equation (9) holds if

$$\sum_k (1 - s_k)[\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F)]Q_{Tk}^\tau \geq 0.$$

This will be satisfied if $\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F) \geq 0$ for all t . This is equivalent to requiring that

$$\sum_{\{i|s_i \geq .5\}} [\gamma(Q_{Ti}^t) - \gamma(Q_{Fi}^t)] \geq 0. \quad (10)$$

Noting that an information service is a stochastic matrix and that the rows add up to one, (10) is satisfied if

$$\sum_{\{i|s_i \leq .5\}} [\gamma(Q_{Fi}^t) - \gamma(Q_{Ti}^t)] \geq 0. \quad (11)$$

Define the cumulative density function of $p(\hat{s}|s_b)$ as

$$G(\hat{s}|s_b) = \sum_{\{a|s_a \leq \hat{s}\}} p(s_a|s_b),$$

where $\hat{s} \in \{s_1, \dots, s_m\} \cup \{s_{\emptyset}\}$. By lemma 1, MLRP holds. This implies that

$$p(\hat{s}|s_k)p(s_j|s_l) > p(s_j|s_k)p(\hat{s}|s_l)$$

for all $s_j < \hat{s}$. Noting that $p(\hat{s}|s_k)p(s_j|s_l) = p(s_j|s_k)p(\hat{s}|s_l)$ when $s_j = \hat{s}$,

$$p(\hat{s}|s_k)p(s_j|s_l) \geq p(s_j|s_k)p(\hat{s}|s_l)$$

for all $s_j \leq \hat{s}$. Summing both sides of this equation from s_0 to $\hat{s}$ with respect to s_j , MLRP implies

$$\frac{p(\hat{s}|s_k)}{p(\hat{s}|s_l)} \geq \frac{G(\hat{s}|s_k)}{G(\hat{s}|s_l)}$$

for all $\hat{s}$. MLRP also implies that

$$p(s_i|s_k)p(\hat{s}|s_l) \geq p(\hat{s}|s_k)p(s_i|s_l)$$

for all $s_i \geq \hat{s}$. Summing both sides of this equation over all $s_i > \hat{s}$, MLRP implies

$$\frac{1 - G(\hat{s}|s_k)}{1 - G(\hat{s}|s_l)} \geq \frac{p(\hat{s}|s_k)}{p(\hat{s}|s_l)}.$$

Combining these two equations we have

$$\frac{1 - G(\hat{s}|s_k)}{1 - G(\hat{s}|s_l)} \geq \frac{G(\hat{s}|s_k)}{G(\hat{s}|s_l)}$$

or

$$\frac{G(\hat{s}|s_l)}{1 - G(\hat{s}|s_l)} \geq \frac{G(\hat{s}|s_k)}{1 - G(\hat{s}|s_k)},$$

which implies $G(\hat{s}|s_l) \geq G(\hat{s}|s_k)$ for any $\hat{s}$ and for signals $s_l < s_k$.

When $s_l = 0$, $G(\hat{s}|0) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Fi}$. Further, when $s_k = 1$, $G(\hat{s}|1) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Ti}$. Thus MLRP implies

$$\sum_{\{i|s_i \leq \hat{s}\}} Q_{Ti}^t \leq \sum_{\{i|s_i \leq \hat{s}\}} Q_{Fi}^t$$

and thus equation (11) holds. The proof for the case where the event is false uses the same logic as the case where the state is true. In this case

$$\mathbb{E}M^V(Q^t|Q^\tau, F) = \sum_k M^V(Q^t|s_k)Q_{Fk}^\tau,$$

and expanding $M^V(Q^t|s_k) = s_k \mathbb{E}V(Q^t|T) + (1 - s_k) \mathbb{E}V(Q^t|F)$,

$$\mathbb{E}M^V(Q^t|Q^\tau, F) - \mathbb{E}V(Q^t|F) = \sum_k [s_k (\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F))] Q_{Fk}^\tau.$$

This is greater or equal to zero if $\mathbb{E}V(Q^t|T) - \mathbb{E}V(Q^t|F) \geq 0$ for all Q^t . We have shown this to be true by MLRP above.

Proof of Lemma 4: Assume that the event is true. The expected probability prediction of forecasters from information service $t \in \{E, N\}$ in state $o \in \{T, F\}$ is given by

$$\mathbb{E}P(Q^t|o) = \sum_{s_i} s_i Q_{oi}^t.$$

The meta-prediction of an individual in group t with signal s_k is

$$M^P(Q^t|s_k) = s_k \mathbb{E}P(Q^t|T) + (1 - s_k) \mathbb{E}P(Q^t|F).$$

The expected meta-prediction of forecasters in information service Q^t made by forecasters with information service Q^τ ($\tau \in \{N, E\}$) when the state is o is given by

$$\mathbb{E}M^P(Q^t|Q^\tau, o) = \sum_k M^P(Q^t|s_k)Q_{ok}^\tau.$$

Aggregating up across novices and experts, the expected probability meta-prediction from information service Q^t given state o is

$$\mathbb{E}M^P(Q^t|o) = \theta \mathbb{E}M^P(Q^t|Q^E, o) + (1 - \theta) \mathbb{E}M^P(Q^t|Q^N, o).$$

In the true state, the probability meta-prediction will underestimate the true probability average if for all $t \in \{E, N\}$,

$$\mathbb{E}P(Q^t|T) \geq \mathbb{E}M^P(Q^t|T). \quad (12)$$

We again allow for equality here to account for cases where (i) all individuals know the state is true or (ii) all individuals receive $s_\emptyset$. In these special cases the probability meta-prediction will be equal to the average probability.

Noting that $\mathbb{E}P(Q^t|T) = \theta\mathbb{E}P(Q^t|T) + (1 - \theta)\mathbb{E}P(Q^t|T)$, equation (12) holds if for $t \in \{E, N\}$ and $\tau \in \{E, N\}$,

$$\mathbb{E}P(Q^t|T) \geq \mathbb{E}M^P(Q^t|Q^\tau, T).$$

Next, recalling that (i)

$$\mathbb{E}M^P(Q^t|Q^\tau, o) = \sum_k M^P(Q^t|s_k)Q_{ok}^\tau,$$

(ii) $M^P(Q^t|s_k) = s_k\mathbb{E}P(Q^t|T) + (1 - s_k)\mathbb{E}P(Q^t|F)$, and (iii) $\sum_k Q_{Tk}^t = 1$, equation (12) holds if

$$\sum_k (1 - s_k)[\mathbb{E}P(Q^t|T) - \mathbb{E}P(Q^t|F)] \geq 0.$$

This will be satisfied if $\mathbb{E}P(Q^t|T) - \mathbb{E}P(Q^t|F) \geq 0$ for all t .

Using the notation from Lemma 3, let $G(\hat{s}|0) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Fi}^t$ and $G(\hat{s}|1) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Ti}^t$ and recall that MLRP implies that for any $\hat{s}$

$$G(\hat{s}|0) \leq G(\hat{s}|1).$$

Thus $G(\hat{s}|1)$ First-order stochastically dominates $G(\hat{s}|0)$. An equivalent definition of stochastic dominance is that for any increasing function $u(\hat{s})$,

$$\sum_i u(s_i)Q_{Ti} \geq \sum_i u(s_i)Q_{Fi}$$

Using $u(\hat{s}) = \hat{s}$, this immediately implies that

$$\sum_i s_i[Q_{Ti} - Q_{Fi}] \geq 0,$$

which is equivalent to $\mathbb{E}P(Q^t|T) - \mathbb{E}P(Q^t|F) \geq 0$. The proof for the case where the event is false uses the same logic as the case where the state is true. In this case we want to prove that $\mathbb{E}P(Q^t|F) \leq \mathbb{E}M^P(Q^t|Q^\tau, F)$. By definition,

$$\mathbb{E}M^P(Q^t|Q^\tau, F) = \sum_k M^P(Q^t|s_k)Q_{Fk}^\tau.$$

Expanding out $M^P(Q^t|s_k)Q_{Fk}^\tau$ and using the same steps as above, $\mathbb{E}P(Q^t|F) \leq \mathbb{E}M^P(Q^t|Q^\tau, F)$ if

$$\sum_k (1 - s_k)[\mathbb{E}P(Q^t|F) - \mathbb{E}P(Q^t|T)] \leq 0.$$

We have shown by MLRP that $\mathbb{E}P(Q^t|F) \leq \mathbb{E}P(Q^t|T)$ and thus the condition holds for all k .

Propositions (1) – (6)

Proof of Proposition 1: In this proof, we show that in the SP algorithm, if (i) forecaster i is better-informed than forecaster j and (ii) the prior is unbiased, then the weight given to forecaster i will be strictly less than the weight given to forecaster j .

To begin, note that when conditions (i) and (ii) hold above, either $\sigma_i < \sigma_j < 0.5$ or $\sigma_i > \sigma_j > 0.5$. Thus $V_i = V_j$. Without loss of generality, we concentrate on the case where $\sigma_i > \sigma_j > 0.5$ so that $V_i = V_j = 1$.

We are interested in the sign of $W_i^{SP}(\sigma_i) - W_j^{SP}(\sigma_j)$. If the sign is positive, then weights are increasing in signal and if it is negative, then weights will be decreasing in signal. Noting that the denominators of

$W_i^{SP}(\sigma_i)$ and $W_j^{SP}(\sigma_j)$ are positive and identical, the sign of $W_i^{SP}(\sigma_i) - W_j^{SP}(\sigma_j)$ will be the same as the sign of $|V(\sigma_i) - M^V(Q|\sigma_i)| - |V(\sigma_j) - M^V(Q|\sigma_j)|$.

As noted in the main text

$$M^V(Q|\sigma_k) = \sigma_k \mathbb{E}V(Q|T) + (1 - \sigma_k) \mathbb{E}V(Q|F).$$

Thus, in the case where $\sigma_i > \sigma_j > 0.5$,

$$|V(\sigma_i) - M^V(Q|\sigma_i)| = 1 - \sigma_i \mathbb{E}V(Q|T) - (1 - \sigma_i) \mathbb{E}V(Q|F)$$

and thus

$$\begin{aligned} |V(\sigma_i) - M^V(Q|\sigma_i)| - |V(\sigma_j) - M^V(Q|\sigma_j)| &= [\sigma_j - \sigma_i] \mathbb{E}V(Q|T) + [(1 - \sigma_j) - (1 - \sigma_i)] \mathbb{E}V(Q|F) \\ &= [\sigma_j - \sigma_i] [\mathbb{E}V(Q|T) - \mathbb{E}V(Q|F)]. \end{aligned}$$

As shown in the proof of Lemma 3, $\mathbb{E}V(Q|T) > \mathbb{E}V(Q|F)$. Thus, since $\sigma_i > \sigma_j$, the sign of $|V(\sigma_i) - M^V(Q|\sigma_i)| - |V(\sigma_j) - M^V(Q|\sigma_j)|$ is negative. Thus, the weights given to forecaster i will be strictly less than the weight given to forecaster j .

Proof of Proposition 2: In this proof, we show that in the SC algorithm, if (i) forecaster i is better-informed than forecaster j , the weight given to i will be strictly more than the weight given to forecaster j .

Consider the case where $\sigma_i > \sigma_j > s_\emptyset$. We are interested in the sign of $W_i^{SC}(\sigma_i) - W_j^{SC}(\sigma_j)$. If the sign is positive, then weights are increasing in signal and if it is negative, then weights will be decreasing in signal. Noting that the denominators of $W_i^{SP}(\sigma_i)$ and $W_j^{SP}(\sigma_j)$ are positive and identical, the sign of $W_i^{SP}(\sigma_i) - W_j^{SP}(\sigma_j)$ will be the same as the sign of $|P(\sigma_i) - M^P(Q|\sigma_i)| - |P(\sigma_j) - M^P(Q|\sigma_j)|$.

As noted in the main text

$$M^P(Q|\sigma_k) = \sigma_k \mathbb{E}P(Q|T) + (1 - \sigma_k) \mathbb{E}P(Q|F)$$

Thus, in the case where $\sigma_i > \sigma_j > 0.5$,

$$|P(\sigma_i) - M^P(Q|\sigma_i)| = \sigma_i - \sigma_i \mathbb{E}P(Q|T) - (1 - \sigma_i) \mathbb{E}P(Q|F)$$

and thus

$$\begin{aligned} |P(\sigma_i) - M^P(Q|\sigma_i)| - |P(\sigma_j) - M^P(Q|\sigma_j)| &= [\sigma_i - \sigma_j] - [\sigma_i - \sigma_j] \mathbb{E}P(Q|T) + [\sigma_i - \sigma_j] \mathbb{E}P(Q|F) \\ &= [\sigma_i - \sigma_j] [1 - (\mathbb{E}P(Q|T) - \mathbb{E}P(Q|F))]. \end{aligned}$$

As shown in the proof of Lemma 4, $0 \leq \mathbb{E}P(Q|F) < \mathbb{E}P(Q|T) \leq 1$. Thus, $[1 - (\mathbb{E}P(Q|T) - \mathbb{E}P(Q|F))]$ is positive. Since, $\sigma_i > \sigma_j$, the sign of $|P(\sigma_i) - M^P(Q|\sigma_i)| - |P(\sigma_j) - M^P(Q|\sigma_j)|$ is positive. As a consequence, the weights given to forecaster i will be strictly greater than the weight given to forecaster j .

Proof of Proposition 3: Define the absolute value of the expected contribution of a forecaster who receives signal s_i in the SP algorithm as:

$$|C^{SP}(Q|s_i)| = \begin{cases} |-M_i^V(s_i)| & \text{if } s_i < 0.5, \\ \frac{1}{2}|-M_i^V(s_i)| + \frac{1}{2}|1 - M_i^V(s_i)| & \text{if } s_i = 0.5, \\ |1 - M_i^V(s_i)| & \text{if } s_i > 0.5. \end{cases}$$

Summing up over forecasters, let $\frac{1}{N} \sum_i^N \mathbb{I}_{\{\sigma_i=s_i\}}$ be the proportion of forecasters that receives signal s_i . Then, Borel's law of large numbers implies that with probability one,

$$\frac{1}{N} \sum_i^N \mathbb{I}_{\{\sigma_i=s_j\}} \rightarrow Q_{oj} \text{ as } N \rightarrow \infty$$

for $o \in \{T, F\}$ and for all s_j . Since each forecaster receives an independent signal, this result implies that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_i^N |V_i - M_i^V| = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{s_j} \left(\sum_i^N |V_i - M_i^V| \mathbb{I}_{\{\sigma_i=s_j\}} \right) = \sum_{s_j} |C^{SP}(Q|s_j)| Q_{oj}$$

and

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_i^N |V_i - M_i^V| V_i = \sum_{\{s_j | s_j \geq 0.5\}} |1 - M^V(Q|s_j)| \gamma(Q_{oj})$$

Combining these two results,

$$\lim_{N \rightarrow \infty} \sum_i^N W_i^{SP} V_i = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_i^N |V_i - M_i^V| V_i}{\frac{1}{N} \sum_i^N |V_i - M_i^V|} = \frac{\sum_{\{s_j | s_j \geq 0.5\}} |1 - M^V(Q|s_j)| \gamma(Q_{oj})}{\sum_{s_j} |C^{SP}(Q|s_j)| Q_{oj}}. \quad (13)$$

Likewise, in the SC algorithm,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_i^N |P_i - M_i^P| = \sum_{s_j} |s_j - M^P(s_j)| Q_{oj}$$

and

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_i^N |P_i - M_i^P| \mathbb{I}_{\{P_i > M_i^P\}} = \sum_{\{s_j | s_j \geq s_{\emptyset}\}} |s_j - M^P(s_j)| Q_{oj}.$$

Combining these results,

$$\lim_{N \rightarrow \infty} \sum_i^N W_i^{SC} \mathbb{I}_{\{P_i > M_i^P\}} = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_i^N |P_i - M_i^P| \mathbb{I}_{\{P_i > M_i^P\}}}{\frac{1}{N} \sum_i^N |P_i - M_i^P|} = \frac{\sum_{\{s_j | s_j \geq s_{\emptyset}\}} |s_j - M^P(s_j)| Q_{oj}}{\sum_{s_j} |s_j - M^P(s_j)| Q_{oj}}. \quad (14)$$

Equations (13) and (14) imply that

$$\lim_{N \rightarrow \infty} \sum_i^N W_i^{SC} \mathbb{I}_{\{P_i > M_i^P\}} \geq \lim_{N \rightarrow \infty} \sum_i^N W_i^{SP} V_i$$

if and only if

$$\frac{\sum_{\{s_j | s_j \geq s_{\emptyset}\}} |s_j - M^P(s_j)| Q_{oj}}{\sum_{s_j} |s_j - M^P(s_j)| Q_{oj}} \geq \frac{\sum_{\{s_j | s_j \geq 0.5\}} |1 - M^V(Q|s_j)| \gamma(Q_{oj})}{\sum_{s_j} |C^{SP}(Q|s_j)| Q_{oj}} \quad (15)$$

for $o \in \{T, F\}$. We will prove this relationship holds in the true state when the problem is unbiased and $s_{\emptyset} = 0.5$. The case for the false state is identical and is omitted.

Starting with the left hand side of equation (15),

$$\sum_{\{s_j | s_j \geq 0.5\}} |s_j - M^P(s_j)| Q_{Tj} = \sum_{\{s_j | s_j \geq 0.5\}} |s_j - s_j \mathbb{E}(P|T) - (1 - s_j) \mathbb{E}(P|F)| Q_{Tj}$$

By the law of total probability $\mathbb{E}(P) = 0.5\mathbb{E}(P|T) + 0.5\mathbb{E}(P|F)$. Noting that $\mathbb{E}(P) = 0.5$ in the unbiased case, this implies $\mathbb{E}(P|F) = 1 - \mathbb{E}(P|T)$. Thus:

$$\sum_{\{s_j | s_j \geq 0.5\}} |s_j - M^P(s_j)|Q_{Tj} = \sum_{\{s_j | s_j \geq 0.5\}} |2s_j - 1|[1 - \mathbb{E}(P|T)]Q_{Tj}.$$

Likewise,

$$\sum_{s_j} |s_j - M^P(s_j)|Q_{Tj} = \sum_{s_j} |2s_j - 1|[1 - \mathbb{E}(P|T)]Q_{Tj}.$$

Thus, the left hand side of equation (15) in the true state with an unbiased prior is equal to:

$$\frac{\sum_{\{s_j | s_j \geq 0.5\}} |2s_j - 1|Q_{Tj}}{\sum_{s_j} |2s_j - 1|Q_{Tj}}. \quad (16)$$

We can also simplify the right hand side of (15). By the law of total probability, $\mathbb{E}(V) = .5\mathbb{E}(V|T) + .5\mathbb{E}(V|F)$, and thus $\mathbb{E}(V|F) = 2\mathbb{E}(V) - \mathbb{E}(V|T)$. Thus, for $s_j \geq 0.5$:

$$\begin{aligned} |1 - M^V(Q|s_j)| &= 1 - s_j\mathbb{E}(V|T) - (1 - s_j)[2\mathbb{E}(V) - \mathbb{E}(V|T)] \\ &= 1 - \mathbb{E}(V) - (2s_j - 1)[\mathbb{E}(V|T) - \mathbb{E}(V)] \\ &= 1 - \mathbb{E}(V) - |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]. \end{aligned}$$

Likewise, for $s_j < 0.5$:

$$\begin{aligned} |-M^V(Q|s_j)| &= |-s_j\mathbb{E}(V|T) - (1 - s_j)[2\mathbb{E}(V) - \mathbb{E}(V|T)]| \\ &= |-\mathbb{E}(V) + (1 - 2s_j)[\mathbb{E}(V|T) - \mathbb{E}(V)]| \\ &= \mathbb{E}(V) - |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]. \end{aligned}$$

Finally, if $s_j = 0.5$, then

$$|C^{SP}(Q|s_j)| = \frac{1}{2}[1 - \mathbb{E}(V)] + \frac{1}{2}\mathbb{E}(V).$$

Noting that when $s_j = 0.5$, $\frac{1}{2}[1 - \mathbb{E}(V)]Q_{Ti} = [1 - \mathbb{E}(V)]\gamma(Q_{Ti})$ and $\frac{1}{2}\mathbb{E}(V)Q_{Ti} = [\mathbb{E}(V)]\gamma(Q_{Ti})$, the right hand side of (15) can be rewritten as

$$\frac{\sum_{\{s_j | s_j \geq .5\}} [1 - E(V)]\gamma(Q_{Tj}) - \sum_{\{s_j | s_j \geq .5\}} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]\gamma(Q_{Tj})}{\sum_{\{s_j | s_j \leq .5\}} [E(V)]\gamma(Q_{Tj}) + \sum_{\{s_j | s_j \geq .5\}} [1 - E(V)]\gamma(Q_{Tj}) + \sum_{s_j \neq 0.5} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]Q_{Tj}}.$$

This is equivalent to

$$\frac{\sum_{\{s_j | s_j \geq .5\}} [1 - E(V)]\gamma(Q_{Tj}) - \sum_{\{s_j | s_j \geq .5\}} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]Q_{Tj}}{\sum_{\{s_j | s_j \leq .5\}} [E(V)]\gamma(Q_{Tj}) + \sum_{\{s_j | s_j \geq .5\}} [1 - E(V)]\gamma(Q_{Tj}) + \sum_{s_j} |2s_j - 1|[\mathbb{E}(V|T) - \mathbb{E}(V)]Q_{Tj}} \quad (17)$$

since $|2s_j - 1| = 0$ for $s_j = 0.5$ and $\gamma(Q_{Tj}) = Q_{Tj}$ for $s_j \neq 0.5$.

Noting that $\sum_{\{s_j | s_j \geq .5\}} [1 - E(V)]\gamma(Q_{Ti}) = [1 - \mathbb{E}(V)]\mathbb{E}(V|T)$, and $\sum_{\{s_j | s_j \leq .5\}} [E(V)]\gamma(Q_{Ti}) = \mathbb{E}(V)[1 - \mathbb{E}(V|T)]$, equation (17) can be rewritten as:

$$\frac{\frac{\mathbb{E}(V|T)[1 - \mathbb{E}(V)]}{\mathbb{E}(V|T) - \mathbb{E}(V)} - \sum_{\{s_j | s_j \geq .5\}} |2s_j - 1|Q_{Tj}}{\frac{[1 - \mathbb{E}(V|T)]\mathbb{E}(V) + \mathbb{E}(V|T)[1 - \mathbb{E}(V)]}{\mathbb{E}(V|T) - \mathbb{E}(V)} - \sum_{s_j} |2s_j - 1|Q_{Tj}}. \quad (18)$$

Cross multiplication shows that for any values of x , y , a , and b with $x > b > 0$ and $y > a > 0$,

$$\frac{a}{b} \geq \frac{x - a}{y - b}$$

if and only if

$$\frac{a}{b} \geq \frac{x}{y}$$

Thus, to show that equation (16) is greater than equation (18), it is sufficient to show that

$$\frac{\sum_{\{s_j | s_j \geq 0.5\}} |2s_j - 1| Q_{Tj}}{\sum_{s_j} |2s_j - 1| Q_{Tj}} \geq \frac{\mathbb{E}(V|T)[1 - \mathbb{E}(V)]}{\mathbb{E}(V|T)[1 - \mathbb{E}(V)] + [1 - \mathbb{E}(V|T)]\mathbb{E}(V)}. \quad (19)$$

Next, note that

$$\begin{aligned} \sum_{\{s_j | s_j \geq 0.5\}} |2s_j - 1| Q_{Tj} &= \sum_{\{s_j | s_j \geq 0.5\}} (2s_j - 1) Q_{Tj} \\ &= 2\mathbb{E}(V|T)[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5] \end{aligned}$$

and

$$\begin{aligned} \sum_{s_j} |2s_j - 1| Q_{Tj} &= \sum_{\{s_j | s_j \geq 0.5\}} (2s_j - 1) Q_{Tj} + \sum_{\{s_j | s_j \leq 0.5\}} (1 - 2s_j) Q_{Tj} \\ &= 2\mathbb{E}(V|T)[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5] + 2[1 - \mathbb{E}(V|T)][.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)] \end{aligned}$$

Thus, we can rewrite the condition in (19) as

$$\frac{\mathbb{E}(V|T)[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5]}{\mathbb{E}(V|T)[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5] + [1 - \mathbb{E}(V|T)][.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)]} \geq \frac{\mathbb{E}(V|T)[1 - \mathbb{E}(V)]}{\mathbb{E}(V|T)[1 - \mathbb{E}(V)] + [1 - \mathbb{E}(V|T)]\mathbb{E}(V)}$$

or, equivalently

$$\frac{\mathbb{E}(V|T) \frac{[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5]}{[.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)]}}{\mathbb{E}(V|T) \frac{[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5]}{[.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)]} + [1 - \mathbb{E}(V|T)]} \geq \frac{\mathbb{E}(V|T) \frac{[1 - \mathbb{E}(V)]}{\mathbb{E}(V)}}{\mathbb{E}(V|T) \frac{[1 - \mathbb{E}(V)]}{\mathbb{E}(V)} + [1 - \mathbb{E}(V|T)]}.$$

Cross multiplication shows that for any x , y , and α with $x \geq 0$, $y \geq 0$, and $\alpha \in (0, 1)$,

$$\frac{\alpha x}{\alpha x + (1 - \alpha)} \geq \frac{\alpha y}{\alpha y + (1 - \alpha)}$$

if and only if $x \geq y$. Thus, equation (19) is satisfied if

$$\frac{[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5]}{.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)} \geq \frac{[1 - \mathbb{E}(V)]}{\mathbb{E}(V)}. \quad (20)$$

As a final step, note that

$$\mathbb{E}(s_j) = \mathbb{E}(s_j | s_j \geq .5) \mathbb{E}(V|T) + \mathbb{E}(s_j | s_j \leq .5) [1 - \mathbb{E}(V|T)].$$

Since the decision problem is unbiased, $\mathbb{E}(s_j) = 0.5$, and thus

$$[\mathbb{E}(s_j | s_j \geq .5) - .5] \mathbb{E}(V|T) = [.5 - \mathbb{E}(s_j | s_j \leq .5)] [1 - \mathbb{E}(V|T)].$$

Rearranging this equation, it is the case that

$$\frac{[\mathbb{E}(s_j | s_j \geq 0.5) - .5]}{.5 - \mathbb{E}(s_j | s_j \leq 0.5)} = \frac{[1 - \mathbb{E}(V)]}{\mathbb{E}(V)}$$

Finally, since $Q_{Tj} > Q_{Fj}$ for all $s_j > 0.5$,

$$\mathbb{E}(s_j | s_j \geq 0.5, T) > \mathbb{E}(s_j | s_j \geq 0.5).$$

Likewise, $Q_{Tj} < Q_{Fj}$ for all $s_j < 0.5$. Thus

$$\mathbb{E}(s_j | s_j \leq 0.5, T) < \mathbb{E}(s_j | s_j \leq 0.5).$$

Thus, it is the case that

$$\frac{[\mathbb{E}(s_j | s_j \geq 0.5, T) - .5]}{0.5 - \mathbb{E}(s_j | s_j \leq 0.5, T)} \geq \frac{[\mathbb{E}(s_j | s_j \geq 0.5) - .5]}{0.5 - \mathbb{E}(s_j | s_j \leq 0.5)} = \frac{[1 - \mathbb{E}(V)]}{\mathbb{E}(V)}.$$

Thus equation (20) is satisfied. This implies that equation (19) is also satisfied and that equation (16) is greater than equation (18).

Proof of Proposition 4: A forecaster with signal s_k will make a probabilistic forecast of s_k . Thus, given an outcome state o , the expected prediction from information service Q^t is given by

$$P(Q^t | o) = \sum_{\{i | s_i \geq 0\}} Q_{oi}^t s_i.$$

Aggregating over both information systems, the expected prediction of the population in state o is given by

$$\mathbb{E}P(\theta | o) := \theta \mathbb{E}P(Q^E | o) + (1 - \theta) \mathbb{E}P(Q^N | o)$$

In the absence of any information service, the probabilistic forecast of each individual would be $s_\emptyset$. By the law of total expectations, the posteriors are a mean-preserving spread of the prior, and thus we have

$$s_\emptyset = s_\emptyset \mathbb{E}P(Q^\tau | T) + (1 - s_\emptyset) \mathbb{E}P(Q^\tau | F).$$

for $\tau \in \{E, N\}$. This also implies that

$$s_\emptyset = s_\emptyset \mathbb{E}P(\theta | T) + (1 - s_\emptyset) \mathbb{E}P(\theta | F)$$

and that

$$\mathbb{E}P(\theta | F) = \frac{s_\emptyset}{1 - s_\emptyset} [1 - \mathbb{E}P(\theta | T)]. \quad (21)$$

A forecaster with signal s_k 's meta-prediction about the others is equal to

$$M^P(\theta | s_k) = s_k \mathbb{E}P(\theta | T) + (1 - s_k) \mathbb{E}P(\theta | F).$$

Substituting in for $\mathbb{E}P(\theta | F)$ using (21), the meta-prediction of an individual with signal s_k can be expressed as

$$M^P(\theta | s_k) = s_k \mathbb{E}P(\theta | T) + (1 - s_k) \frac{s_\emptyset}{1 - s_\emptyset} [1 - \mathbb{E}P(\theta | T)].$$

The total contribution of an individual is based on the difference between the individual's prediction and meta-prediction. For an individual with signal s_k ,

$$s_k - M^P(\theta | s_k) = s_k - s_k \mathbb{E}P(\theta | T) - (1 - s_k) \frac{s_\emptyset}{1 - s_\emptyset} [1 - \mathbb{E}P(\theta | T)]$$

or, equivalently:

$$s_k - M^P(\theta | s_k) = \frac{s_k - s_\emptyset}{1 - s_\emptyset} [1 - \mathbb{E}P(\theta | T)]. \quad (22)$$

Note first that the difference between an individual's signal and his or her meta-prediction is zero at $s_\emptyset$ and is linearly increasing in s_k . This feature implies that the weight of an individual with signal s_k , proportional to $|s_k - M^P(\theta|s_k)|$, is directly related to the informativeness of the posterior that an individual holds relative to the prior. Thus, individuals with more informative posteriors (an ex-post notion of expertise) will be weighted proportionally more than individuals with less informative posteriors.

We now show that $\mathbb{E}[P(Q^E|T)] - \mathbb{E}[M^P(\theta|Q^E, T)] \geq \mathbb{E}[P(Q^N|T)] - \mathbb{E}[M^P(\theta|Q^N, T)]$. First note that because $\sum_i Q_{Ti}^E = 1$

$$\begin{aligned} \mathbb{E}[P(Q^E|T)] - \mathbb{E}[M^P(\theta|Q^E, T)] &= \sum_i \left[\frac{s_i - s_\emptyset}{1 - s_\emptyset} [1 - \mathbb{E}P(\theta|T)] \right] Q_{Ti}^E \\ &= \frac{[1 - \mathbb{E}P(\theta|T)]}{1 - s_\emptyset} [(\sum_i s_i Q_{Ti}^E) - s_\emptyset]. \end{aligned}$$

Thus, $\mathbb{E}[P(Q^E|T)] - \mathbb{E}[M^P(\theta|Q^E, T)] - \mathbb{E}[P(Q^N|T)] + \mathbb{E}[M^P(\theta|Q^N, T)]$ is equal to

$$\frac{[1 - \mathbb{E}P(\theta|T)]}{1 - s_\emptyset} [(\sum_i s_i Q_{Ti}^E) - (\sum_i s_i Q_{Ti}^N)]$$

The sign of this equation will be positive if

$$\sum_i s_i Q_{Ti}^E \geq \sum_i s_i Q_{Ti}^N,$$

or, equivalently if $\mathbb{E}[s_i|Q^E, T] \geq \mathbb{E}[s_i|Q^N, T]$.

We now show that $\mathbb{E}[s_i|Q^E, T] \geq \mathbb{E}[s_i|Q^N, T]$. To do so, we will use Blackwell's Theorem Blackwell (1951):

Blackwell's Theorem *For information service Q^E to be more informative than Q^N it is necessary and sufficient that the value of information in service Q^E is greater than the value of information in service Q^N for all sets of terminal actions, all utility functions, and all prior beliefs.*

By Assumption 1, Q^E is more informative than Q^N . Let the action set $V \in \{T, F\}$ correspond to voting on whether an answer is true or false, and consider a utility function $U(V, o)$ that maps actions and states of the world into outcomes. Let $U(T, T) = 1$, $U(F, F) = 0$, $U(F, T) = 0$, and $U(T, F) = 0$. Given a signal s_i , expected utility is maximized by choosing $a = T$ in all states. The expected utility of this strategy given signal s_i is

$$\mathbb{E}[U(Q^t|s_i)] = U(T, T)s_i = s_i$$

By Blackwell's theorem, the expected utility of information service Q^E is higher than the expected value of information service Q^N for any utility function and any prior belief. Using an initial prior of $P(T) = 1$,

$$\mathbb{E}[U(Q^t)] = \sum_i \mathbb{E}[U(Q^t|s_i)] Q_{Ti}^t = \sum_i s_i Q_{Ti}^t$$

Thus, $\mathbb{E}[U(Q^E)] > \mathbb{E}[U(Q^N)]$ immediately implies

$$\sum_i s_i Q_{Ti}^E > \sum_i s_i Q_{Ti}^N,$$

which implies that the sign of $\sum_i s_i Q_{Ti}^E - \sum_i s_i Q_{Ti}^N$ is positive.

The proof for the False state has an identical structure to the proof used for the True state. $\mathbb{E}[M^P(\theta|Q^E, F)] - \mathbb{E}[P(Q^E|F)] - \mathbb{E}[M^P(\theta|Q^N, T)] - \mathbb{E}[P(Q^N|T)]$ is equal to

$$\frac{[1 - \mathbb{E}P(\theta|T)]}{1 - s_{\emptyset}} [(\sum_i s_i Q_{Fi}^N) - (\sum_i s_i Q_{Fi}^E)]$$

The sign of this equation will be positive if

$$\sum_i -s_i Q_{Fi}^E > \sum_i -s_i Q_{Fi}^N.$$

or, equivalently,

$$\sum_i (1 - s_i) Q_{Fi}^E > \sum_i (1 - s_i) Q_{Fi}^N.$$

The left hand side of this last equation is $\mathbb{E}[1 - s_i|Q^E, F]$ while the right hand side is $\mathbb{E}[1 - s_i|Q^N, T]$. Using Blackwell's theorem with $U(F, F) = 1$, $U(T, F) = U(F, T) = U(T, T) = 0$ and $P(T) = 0$ immediately shows that this condition holds.

Proof of Proposition 5: Assume that the event is true. The expected confidence prediction of forecasters from information service $t \in \{E, N\}$ in state $o \in \{T, F\}$ is given by

$$\mathbb{E}C(Q^t|o) = \sum_i \left(\sum_k c(s_k) R_{ik} \right) Q_{oi}^t.$$

By assumption, all forecasters' probability meta-predictions are fully adaptive. Thus, the (confidence adjusted) meta-prediction of an individual in group t with signal s_k is

$$M^C(Q^t|c(s_k)) = c(s_k) \mathbb{E}C(Q^t|T) + (1 - c(s_k)) \mathbb{E}C(Q^t|F).$$

The expected meta-prediction of forecasters in information service Q^t made by forecasters with information service Q^τ ($\tau \in \{N, E\}$) when the state is o is given by

$$\mathbb{E}M^C(Q^t|Q^\tau, o) = \sum_i \left(\sum_k M^C(Q^t|c(s_k)) R_{ik} \right) Q_{oi}^\tau.$$

Aggregating up across novices and experts, the expected probability meta-prediction from information service Q^t given state o is

$$\mathbb{E}M^C(Q^t|o) = \theta \mathbb{E}M^C(Q^t|Q^E, o) + (1 - \theta) \mathbb{E}M^C(Q^t|Q^N, o).$$

In the true state, the probability meta-prediction will underestimate the true probability average if for all $t \in \{E, N\}$,

$$\mathbb{E}C(Q^t|T) \geq \mathbb{E}M^C(Q^t|T). \quad (23)$$

We allow for equality here to account for cases where (i) all individuals know the state is true or (ii) all individuals receive $s_{\emptyset}$. In these special cases the probability meta-prediction will be equal to the average probability.

Noting that $\mathbb{E}C(Q^t|T) = \theta \mathbb{E}C(Q^t|T) + (1 - \theta) \mathbb{E}C(Q^t|T)$, equation (23) holds if for $t \in \{E, N\}$ and $\tau \in \{E, N\}$,

$$\mathbb{E}C(Q^t|T) \geq \mathbb{E}M^C(Q^t|Q^\tau, T).$$

Next, recalling that (i)

$$\mathbb{E}M^C(Q^t|Q^\tau, o) = \sum_i \left(\sum_k M^C(Q^t|c(s_k))R_{ik} \right) Q_{oi}^\tau,$$

(ii) $M^C(Q^t|c(s_k)) = c(s_k)\mathbb{E}C(Q^t|T) + (1 - c(s_k))\mathbb{E}C(Q^t|F)$, and (iii) $\sum_i (\sum_k R_{ik})Q_{Ti}^\tau = 1$, equation (23) holds if

$$\sum_i \left(\sum_k (1 - c(s_k))[\mathbb{E}C(Q^t|T) - \mathbb{E}C(Q^t|F)]R_{ik} \right) Q_{Ti}^\tau \geq 0.$$

This will be satisfied if $\mathbb{E}C(Q^t|T) - \mathbb{E}C(Q^t|F) \geq 0$ for all t .

Noting that by part (iv) of the definition of systematically miscalibrated, $\sum_k c(s_k)R_{ik} = c(s_i)$, and thus

$$\mathbb{E}C(Q^t|o) = \sum_i c(s_i)Q_{oi}^t.$$

Thus, we need to show that $\sum_i c(s_i)Q_{Ti}^t > \sum_i c(s_i)Q_{Fi}^t$. Using the notation from Lemma 3, let $G(\hat{s}|0) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Fi}^t$ and $G(\hat{s}|1) = \sum_{\{i|s_i \leq \hat{s}\}} Q_{Ti}^t$ and recall that MLRP implies that for any $\hat{s}$

$$G(\hat{s}|0) \leq G(\hat{s}|1).$$

Thus $G(\hat{s}|1)$ first-order stochastically dominates $G(\hat{s}|0)$. An equivalent definition of stochastic dominance is that for any increasing function $u(\hat{s})$,

$$\sum_i u(s_i)Q_{Ti} \geq \sum_i u(s_i)Q_{Fi}$$

Using $u(\hat{s}) = c(\hat{s})$, this immediately implies that

$$\sum_i c(s_i)[Q_{Ti} - Q_{Fi}] \geq 0,$$

which is equivalent to $\mathbb{E}C(Q^t|T) - \mathbb{E}C(Q^t|F) \geq 0$.

The proof for the case where the event is false uses the same logic as the case where the state is true. In this case we want to prove that $\mathbb{E}C(Q^t|F) \leq \mathbb{E}M^C(Q^t|Q^\tau, F)$. By definition,

$$\mathbb{E}M^C(Q^t|Q^\tau, F) = \sum_i \left(\sum_k M^C(Q^t|c(s_k))R_{ik} \right) Q_{Fi}^\tau.$$

Expanding out $M^C(Q^t|s_k)Q_{Fi}^\tau$ and using the same steps as above, $\mathbb{E}C(Q^t|F) \leq \mathbb{E}M^C(Q^t|Q^\tau, F)$ if

$$\sum_i \left(\sum_k (1 - c(s_k))[\mathbb{E}C(Q^t|F) - \mathbb{E}C(Q^t|T)]R_{ik} \right) Q_{Fi}^\tau \leq 0.$$

We have shown by MLRP that $\mathbb{E}C(Q^t|F) \leq \mathbb{E}C(Q^t|T)$ and thus the condition holds for all t .

Proof of Proposition 6: Let $\rho_k^\tau := Q_{Fk}^\tau + Q_{Tk}^\tau$. Then, by the assumption that $s_\emptyset = .5$, Bayes Rule implies

$$Q_{Tk}^\tau = s_k \rho_k^\tau$$

and, by definition,

$$\mathbb{E}V(Q^\tau|T) = \frac{1}{2}Q_{T\emptyset}^\tau + \sum_{\{k|s_k > .5\}} Q_{Tk}^\tau = \frac{1}{2}s_\emptyset \rho_\emptyset^\tau + \sum_{\{k|s_k > .5\}} s_k \rho_k^\tau.$$

Recall that $M^V(\theta|s_k)$ is defined as the probabilistic meta-prediction of a forecaster with signal s_k . Note that $M^V(\theta|s_k)$ is a weighted average of $M^V(Q^E|s_k)$ and $M^V(Q^N|s_k)$:

$$M^V(\theta|s_k) := \theta M^V(Q^E|s_k) + (1 - \theta) M^V(Q^N|s_k).$$

Then, by definition,

$$\mathbb{E}M^V(\theta|Q^\tau, T) = \sum_k M^V(\theta|s_k) Q_{T_k}^\tau = \sum_{\{k|s_k < .5\}} M^V(\theta|s_k) s_k \rho_k^\tau + \frac{1}{2} s_\emptyset \rho_\emptyset^\tau + \sum_{\{k|s_k > .5\}} M^V(\theta|s_k) s_k \rho_k^\tau.$$

By the symmetry assumption, for all $k \leq .5m$, (i) $s_k = 1 - s_{m+2-k}$, (ii) $M^V(\theta|s_k) = M^V(\theta|1 - s_{m+2-k})$, and (iii) $\rho_k^\tau = \rho_{m+2-k}^\tau$. This implies

$$\begin{aligned} \sum_{\{k|s_k < .5\}} M^V(\theta|s_k) s_k \rho_k^\tau &= \sum_{\{k|s_{m+2-k} > .5\}} M^V(\theta|s_k) s_k \rho_k^\tau \\ &= \sum_{\{k|s_{m+2-k} > .5\}} M^V(\theta|1 - s_{m+2-k}) (1 - s_{m+2-k}) \rho_{m+2-k}^\tau \\ &= \sum_{\{l|s_l > .5\}} M^V(\theta|1 - s_l) (1 - s_l) \rho_l^\tau, \end{aligned}$$

where $l = m + 2 - k$. Noting that $l \in \{.5m + 2, \dots, m + 1\}$ and shares the same indexes as the original set $\{k|s_k > .5\}$, we can combine terms and rewrite

$$\mathbb{E}M^V(\theta|Q^\tau, T) = \frac{1}{2} s_\emptyset \rho_\emptyset^\tau + \sum_{\{k|s_k > .5\}} [M^V(\theta|s_k) s_k + M^V(\theta|1 - s_k) (1 - s_k)] \rho_k^\tau.$$

Using this representation of the meta prediction, the expected total contribution of an individual in group τ is:

$$\mathbb{E}V(Q^\tau|T) - \mathbb{E}M^V(\theta|Q^\tau, T) = \sum_{\{k|s_k > .5\}} [s_k - s_k M^V(\theta|s_k) - (1 - s_k) M^V(\theta|1 - s_k)] \rho_k^\tau. \quad (24)$$

By the definition of $M^V(\theta|s_k)$:

$$\begin{aligned} s_k M^V(\theta|s_k) + (1 - s_k) M^V(\theta|1 - s_k) &= \theta [s_k M^V(Q^E|s_k) + (1 - s_k) M^V(Q^E|1 - s_k)] \\ &\quad + (1 - \theta) [s_k M^V(Q^N|s_k) + (1 - s_k) M^V(Q^N|1 - s_k)]. \end{aligned} \quad (25)$$

Symmetry implies that $\mathbb{E}V(Q^\tau|F) = 1 - \mathbb{E}V(Q^\tau|T)$. Thus

$$\begin{aligned} M^V(Q^\tau|s_k) &= s_k \mathbb{E}V(Q^\tau|T) + (1 - s_k) \mathbb{E}V(Q^\tau|F) \\ &= s_k \mathbb{E}V(Q^\tau|T) + (1 - s_k) (1 - \mathbb{E}V(Q^\tau|T)). \end{aligned}$$

This implies that for $s_k \geq .5$:

$$\begin{aligned} s_k M^V(Q^\tau|s_k) &= s_k [s_k \mathbb{E}V(Q^\tau|T) + (1 - s_k) (1 - \mathbb{E}V(Q^\tau|T))] \\ &= s_k (1 - s_k) + (s_k^2 - s_k (1 - s_k)) \mathbb{E}V(Q^\tau|T) \end{aligned}$$

and

$$\begin{aligned} (1 - s_k) M^V(Q^\tau|1 - s_k) &= (1 - s_k) [(1 - s_k) \mathbb{E}V(Q^\tau|T) + s_k (1 - \mathbb{E}V(Q^\tau|T))] \\ &= s_k (1 - s_k) + ((1 - s_k)^2 - s_k (1 - s_k)) \mathbb{E}V(Q^\tau|T) \end{aligned}$$

Substitution these two simplifications into (25) implies:

$$s_k M^V(\theta|s_k) + (1 - s_k) M^V(\theta|1 - s_k) = 2s_k(1 - s_k) + (2s_k - 1)^2 [\theta \mathbb{E}V(Q^E|T) + (1 - \theta) \mathbb{E}V(Q^N|T)] \quad (26)$$

Let $\mathbb{E}V(\theta|T) := \theta \mathbb{E}V(Q^E|T) + (1 - \theta) \mathbb{E}V(Q^N|T)$ be the expected vote in the true state and note that this quantity is a constant. Then, using the simplification in (26), equation (24) implies

$$\begin{aligned} \mathbb{E}V(Q^\tau|T) - \mathbb{E}M^V(\theta|Q^\tau, T) &= \sum_{\{k|s_k > .5\}} [s_k - 2s_k(1 - s_k) - (2s_k - 1)^2 \mathbb{E}V(\theta|T)] \rho_k^\tau \\ &= \sum_{\{k|s_k > .5\}} [(2s_k - 1)(s_k(1 - \mathbb{E}V(\theta|T)) + (1 - s_k) \mathbb{E}V(\theta|T))] \rho_k^\tau \\ &= \left[\sum_{\{k|s_k > .5\}} \phi(s_k) \rho_k^\tau \right] + \phi(s_\emptyset) Q_{T\emptyset}^\tau, \end{aligned}$$

where $\phi(s_k) = (2s_k - 1)(s_k(1 - \mathbb{E}V(\theta|T)) + (1 - s_k) \mathbb{E}V(\theta|T))$ and $\phi(s_\emptyset) = 0$. Note that if a symmetric information service has only two posteriors that occur with positive probability, s_k and $(1 - s_k)$, $\phi(s_k)$ is the expected difference between an individual's expected vote and their meta-prediction in the true state. This implies that when an information system is symmetric, the total contribution of a forecaster with information system Q^τ is the weighted average of simpler symmetric information systems that contain only two posteriors.

To show that the expected total contribution of an expert is greater or equal to the expected total contribution of a novice, we need to show that

$$(\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E, T)) - (\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T)) \geq 0$$

or, equivalently,

$$\left[\sum_{\{k|s_k > .5\}} \phi(s_k) \rho_k^E \right] - \left[\sum_{\{l|s_l > .5\}} \phi(s_l) \rho_l^N \right] + \phi(s_\emptyset) [Q_{T\emptyset}^E - Q_{T\emptyset}^N] \geq 0$$

We do this in two steps. First, we construct a set of non-negative weights $w_{k,l}$ where (i) $w_{k,l} = 0$ in cases where $l > k$ and (ii)

$$\begin{aligned} \left[\sum_{\{k|s_k > .5\}} \phi(s_k) \rho_k^E \right] - \left[\sum_{\{l|s_l > .5\}} \phi(s_l) \rho_l^N \right] \\ + \phi(s_\emptyset) [Q_{T\emptyset}^E - Q_{T\emptyset}^N] = \sum_{\{k|s_k \geq .5\}} \sum_{\{l|s_l \geq .5\}} [\phi(s_k) - \phi(s_l)] w_{k,l}. \end{aligned} \quad (27)$$

We then show that $\phi(s_k) - \phi(s_l) \geq 0$ for all $k > l$. This guarantees that each element in the RHS of (27) is positive or zero.

The assumption of strict garbling implies that

$$\sum_{\{i|s_i \leq \hat{s}\}} [Q_{Ti}^E + Q_{T(m+2-i)}^E] \geq \sum_{\{i|s_i \leq \hat{s}\}} [Q_{Ti}^N + Q_{T(m+2-i)}^N]$$

for all $\hat{s} \leq s_\emptyset$. Noting that $Q_{T(m+2-i)}^t = Q_{Fi}^t$, strict garbling implies

$$\sum_i \rho_i^E \geq \sum_i \rho_i^N$$

for all $i \in \{1, \dots, \frac{m}{2} + 1\}$. We use this relationship to construct a matrix of weights $W = [w_{ij}]_{(m+1) \times (m+1)}$ where (27) is satisfied.

We begin by constructing a submatrix consisting of the first $(\frac{m}{2} + 1) \times (\frac{m}{2} + 1)$ elements of W . Let

$$V^E = [\rho_1^E, \rho_2^E, \dots, \rho_{\frac{m}{2}}^E, \frac{1}{2}\rho_{\emptyset}^E]$$

be a $\frac{m}{2} + 1$ element vector. Note that $\frac{1}{2}\rho_{\emptyset}^E = Q_{T\emptyset}$ and thus, by construction, the elements of the vector sum to 1. Likewise, define

$$V^{N,1} = [\rho_1^N, \rho_2^N, \dots, \rho_{\frac{m}{2}}^N, \frac{1}{2}\rho_{\emptyset}^N]$$

and note that the sum of these elements add up to 1.

We construct the first row of weights iteratively. For each $j \in \{1, \dots, \frac{m}{2} + 1\}$, let

$$w_{1,j} = \begin{cases} V_j^{N,1} & \text{if } V_1^E - \sum_{k=1}^{j-1} w_{1,k} \geq V_j^{N,1}, \\ V_1^E - \sum_{k=1}^{j-1} w_{1,k} & \text{otherwise.} \end{cases}$$

By the assumption of strict garbling $\rho_1^E \geq \rho_1^N$, and $w_{1,1} = \rho_1^N$. All other weights in the first row are either zero or positive with $V_1^E = \sum_{j=1}^{\frac{m}{2}+1} w_{1,j}$.

We now construct the rest of the weights row by row in an iterative process. At each step $i \in \{2, \dots, \frac{m}{2} + 1\}$, let

$$V^{N,i} = \left[\left(V_1^{N,1} - \sum_{k=1}^{i-1} w_{k,1} \right), \left(V_2^{N,1} - \sum_{k=1}^{i-1} w_{k,2} \right), \dots, \left(V_{\frac{m}{2}+1}^{N,1} - \sum_{k=1}^{i-1} w_{k,\frac{m}{2}+1} \right) \right].$$

Iterating over $j \in \{1, \frac{m}{2} + 1\}$, let

$$w_{i,j} = \begin{cases} V_j^{N,i} & \text{if } V_i^E - \sum_{k=1}^{j-1} w_{i,k} \geq V_j^{N,i}, \\ V_i^E - \sum_{k=1}^{j-1} w_{i,k} & \text{otherwise.} \end{cases}$$

By the assumption of strict garbling, $\sum_{k=1}^j w_{k,j} = V_j^{N,1}$. Thus, by the construction of the vector $V^{N,i}$, $w_{i,j} = 0$ for all $i > j$. Combined, these two conditions imply $\sum_i w_{i,j} = V_j^{N,1}$ for all j in the submatrix. Further, since both vectors V^E and $V^{N,1}$ sum to 1 by construction, $\sum_j w_{i,j} = V_i^E$ for all i .

Taken together, the construction of the submatrix generates a set of weights such that we can recover the elements of $V^{N,1}$ by adding the elements of the column together. As the first $\frac{m}{2}$ elements of $V^{N,1}$ correspond to $\{\rho_1^N, \dots, \rho_{\frac{m}{2}}^N\}$, we can relate the weight matrix to ρ_j^N by adding the elements of column j together. Likewise, we can recover elements of V^E by adding the elements of the rows together. As the first $\frac{m}{2}$ elements of V^E correspond to $\{\rho_1^E, \dots, \rho_{\frac{m}{2}}^E\}$, we can relate the weight matrix to ρ_i^E by adding the elements of row i together.

We now take advantage of symmetry to construct the weights for elements of W where $i \in \{\frac{m}{2} + 1, \dots, m + 1\}$ and $j \in \{\frac{m}{2} + 1, \dots, m + 1\}$. To avoid confusion with the previous step, let $k \in \{\frac{m}{2} + 1, \dots, m + 1\}$ represent the rows in this submatrix of W and let $l \in \{\frac{m}{2} + 1, \dots, m + 1\}$ represent columns. Next, let $w_{k,l} = w_{(m+2-k), (m+2-l)}$. Note that by reflection, $w_{k,l} = 0$ if $l > k$. All other weights are greater or equal to zero.

By symmetry, $\rho_k^E = \rho_{m+2-k}^E$. Thus for all $k \in \{\frac{m}{2} + 2, \dots, m + 1\}$, $\rho_k^E = V_{m+2-k}^E$ and

$$\phi(s_k)\rho_k^E = \sum_{\{l | s_l \geq .5\}} \phi(s_k)w_{k,l}$$

Likewise, if $k = \frac{m}{2} + 1$, $Q_{T\emptyset}^E = V_{m+2-k}^E$ and

$$\phi(s_{\emptyset})Q_{T\emptyset}^E = \sum_{\{l | s_l \geq .5\}} \phi(s_{\emptyset})w_{k,l}.$$

This implies

$$\left[\sum_{\{k|s_k > .5\}} \phi(s_k) \rho_k^E \right] + \phi(s_\emptyset) Q_{T\emptyset}^E = \sum_{\{k|s_k \geq .5\}} \sum_{\{l|s_l \geq .5\}} \phi(s_k) w_{k,l} \quad (28)$$

Using the same logic,

$$\phi(s_l) \rho_l^N = \sum_{\{k|s_k \geq .5\}} \phi(s_l) w_{k,l}$$

for all $l \in \{\frac{m}{2} + 2, \dots, m + 1\}$ and

$$\phi(s_\emptyset) Q_{T\emptyset}^N = \sum_{\{k|s_k \geq .5\}} \phi(s_\emptyset) w_{k,l}.$$

when $l = \frac{m}{2} + 1$. Thus

$$\left[\sum_{\{l|s_l > .5\}} \phi(s_l) \rho_l^N \right] + \phi(s_\emptyset) Q_{T\emptyset}^N = \sum_{\{k|s_k \geq .5\}} \sum_{\{l|s_l \geq .5\}} \phi(s_l) w_{k,l} \quad (29)$$

Subtracting (29) from (28) implies that (27) holds.

By Assumption 6, $\mathbb{E}V(\theta|T) < .75$. We now show that when $\mathbb{E}V(\theta|T) < .75$, $\phi(s_k) > \phi(s_l)$ if $s_k > s_l \geq .5$. By definition

$$\begin{aligned} \phi(s_k) - \phi(s_l) &= (2s_k - 1)[s_k(1 - \mathbb{E}V(\theta|T)) + (1 - s_k)\mathbb{E}V(\theta|T)] \\ &\quad - (2s_l - 1)[s_l(1 - \mathbb{E}V(\theta|T)) + (1 - s_l)\mathbb{E}V(\theta|T)] \\ &= 2(s_k - s_l)[s_k(1 - \mathbb{E}V(\theta|T)) + (1 - s_k)\mathbb{E}V(\theta|T)] \\ &\quad - (2s_l - 1)[(s_k - s_l)(2\mathbb{E}V(\theta|T) - 1)] \\ &= (s_k - s_l)[2s_k - 4s_k\mathbb{E}V(\theta|T) + 2V(\theta|T) - (2s_l - 1)(2\mathbb{E}V(\theta|T) - 1)] \end{aligned}$$

We have assumed that $s_k > s_l$. This implies that $(s_k - s_l)$ is strictly positive and $\phi(s_k) > \phi(s_l)$ if and only

$$2s_k - 4s_k\mathbb{E}V(\theta|T) + 2\mathbb{E}V(\theta|T) - (2s_l - 1)(2\mathbb{E}V(\theta|T) - 1) > 0. \quad (30)$$

Notice that (30) is decreasing in s_l . This implies that:

$$\begin{aligned} 2s_k - 4s_k\mathbb{E}V(\theta|T) + 2\mathbb{E}V(\theta|T) - (2s_l - 1)(2\mathbb{E}V(\theta|T) - 1) &> 2s_k - 4s_k\mathbb{E}V(\theta|T) + 2\mathbb{E}V(\theta|T) \\ &\quad - (2s_k - 1)(2\mathbb{E}V(\theta|T) - 1). \end{aligned}$$

Thus, a sufficient condition for $\phi(s_k) - \phi(s_l)$ to be positive is for

$$2s_k - 4s_k\mathbb{E}V(\theta|T) + 2\mathbb{E}V(\theta|T) - (2s_k - 1)(2\mathbb{E}V(\theta|T) - 1) \geq 0.$$

Rearranging this equation, $\phi(s_k) - \phi(s_l)$ is positive if

$$\frac{\mathbb{E}V(\theta|T) - .25}{2\mathbb{E}V(\theta|T) - 1} \geq s_k.$$

Further, noting that $s_k \in (.5, 1]$, $\phi(s_k) - \phi(s_l)$ is positive if

$$\frac{\mathbb{E}V(\theta|T) - .25}{2\mathbb{E}V(\theta|T) - 1} \geq 1.$$

The LHS is decreasing in $\mathbb{E}V(\theta|T)$ and equal to one when $\mathbb{E}V(\theta|T) = .75$. Thus $\phi(s_k) > \phi(s_l)$ whenever $\mathbb{E}V(\theta|T) < .75$

By the construction of the weight matrix, there exists at least one element $w_{k,l}$ with $k > l$ where $w_{k,l} > 0$. For this element $[\phi(s_k) - \phi(s_l)]w_{k,l} > 0$. Noting that $w_{k,l} = 0$ when $k < l$, it follows that all other elements of (27) are either positive or zero and thus

$$(\mathbb{E}V(Q^E|T) - \mathbb{E}M^V(\theta|Q^E, T)) - (\mathbb{E}V(Q^N|T) - \mathbb{E}M^V(\theta|Q^N, T))$$

is positive.

Online Appendix D: Additional Empirical Results

D1. Estimated weights in Experiment 2

In this appendix, we estimate the weights of individual forecasters in Experiment 2. Our analysis is identical to that of Experiment 1 (see Section 3.1.2). For the SC algorithm, we once again estimate the prior by regressing the probability meta-prediction on the probability forecast and finding where this line crosses the identity line. Over this entire dataset for Experiment 2, the prior is estimated to be at .68. Thus, the data in this dataset is also biased towards true.

Figure 9 shows the estimated weights in the SP algorithm (top panel) and SC algorithm (bottom panel) as a function of the signal they received for all five grade levels combined. As before, the black solid line in each graph shows the predictions from each theoretical model while the dashed line shows the estimates from a non-parametric kernel regression. As seen in the top graph, both the magnitude of forecasters' signals ($|P_{i,k} - 0.5|$) and their votes ($V_{i,k}$) were significant negative predictors of their weight in the SP algorithm, $\beta_1 = -0.70$, $F(1, 458) = 2108.6$, $p < .001$ and $\beta_2 = -0.06$, $F(1, 458) = 177.0$, $p < .001$. There is once again a gap in the weight function at 0.5 in the same direction as before, which indicates that there is bias in the dataset towards answering true. However, unlike the states data, the gap is much smaller, and individuals who are certain that an event is true or false have weights that are less than half that of an individual who has a vote meta-prediction of 0.5.

As seen in the bottom panel, the SC algorithm has weights that are increasing in the distance away from the predicted prior, with a significant and large positive slope in the model specification that is consistent with the theoretical predictions, $\beta_1 = 0.22$, $F(1, 458) = 118.3$, $p < .001$. Better-informed forecasters are therefore generating larger weights in the SC algorithm than lesser-informed forecasters.

Our results here are therefore consistent with our theoretical model predictions: weights in the SP algorithm are decreasing in the distance from the 0.5 in the quiz dataset whereas weights in the SC algorithm are increasing in the distance away from the estimated uninformed prior.

D2. Weights and Expertise in Experiment 1 and 2

In the main text, we divided forecasters into high-performers and low-performers as a way of separating forecasters who are likely to be experts from those who are likely to be novices. In this section, we study an alternative specification where we further subdivide forecasters into quartiles to better understand how forecasters with different track-records contribute to the algorithm.

Similar to our approach in the main text, we sorted forecasters based on their mean accuracy on all other decision problems in the test set using leave-one-out cross-validation. Next, we divided forecasters into four quartiles containing equal numbers of forecasters and examined the average weight assigned by each algorithm to each quartile of forecasters over the test set. By construction, the weights in the four quartiles add up to one.

Figure 10 shows the alternative quartile specification for Experiment 1. As seen in the right hand panel of Figure 10, both the SP and SC algorithm over-weight forecasters in the highest two quartiles and under-weight forecasters in the lowest two quartiles on false questions. However, the SC algorithm assigns substantially

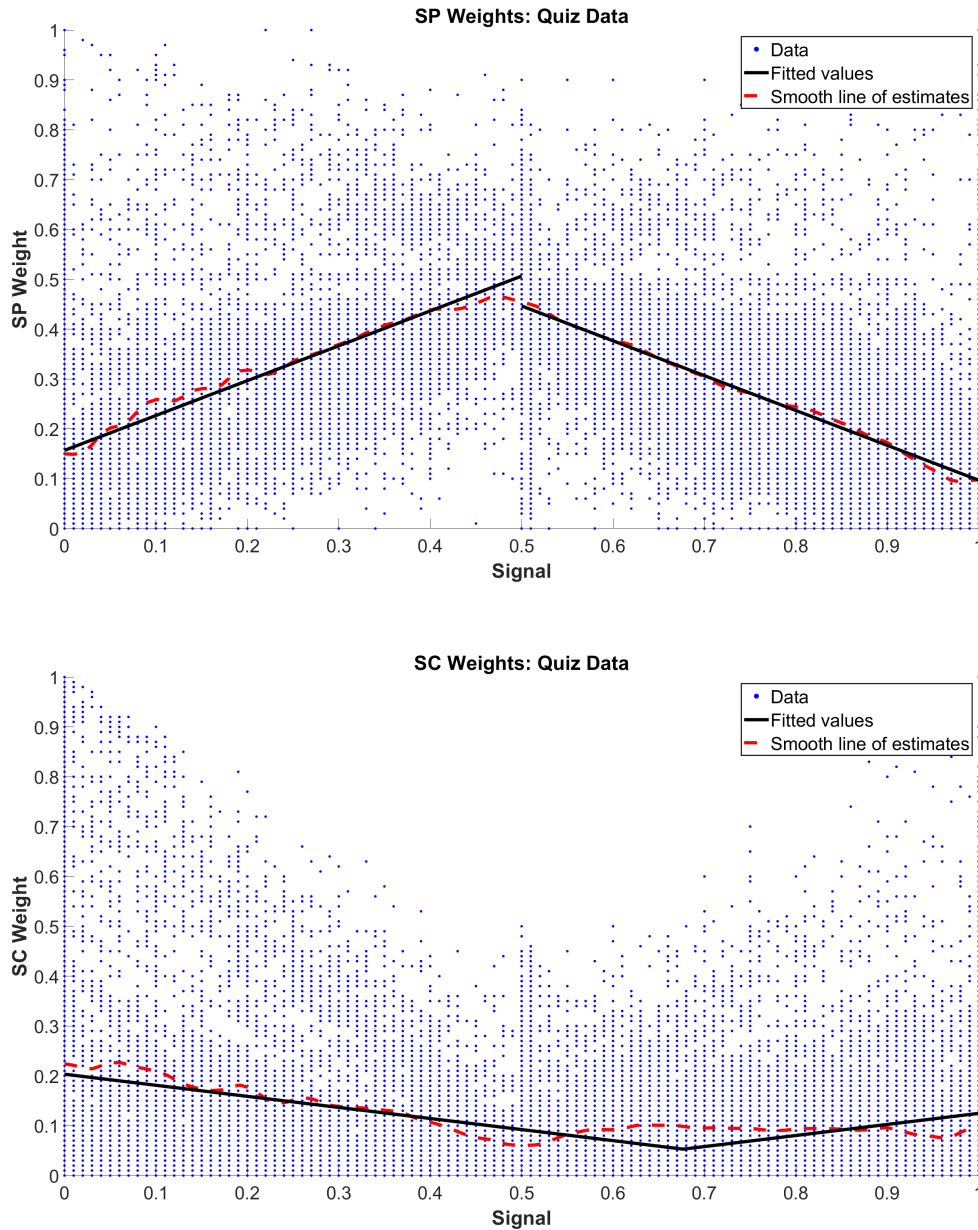


Figure 9 The relationship between forecasters' posterior and the weight assigned to them by the SP algorithm (top panel) and the SC algorithm (bottom panel) for the Quiz Data. The solid black lines are the predictions from the theoretical models. The dashed line is from a non-parametric kernel regression.

more weight to forecasters in the highest quartile and substantially less weight to forecasters in the lowest quartile, compared to the SP algorithm. As seen in the left panel, the SC also over-weights forecasters in the highest quartile and under-weights forecasters in the lowest quartile in true questions, while the weights in the SP algorithm are similar across the four quartiles. Thus, the SC algorithm appears to be more effective than the SP algorithm at assigning weight to forecasters who are correct more often on average.

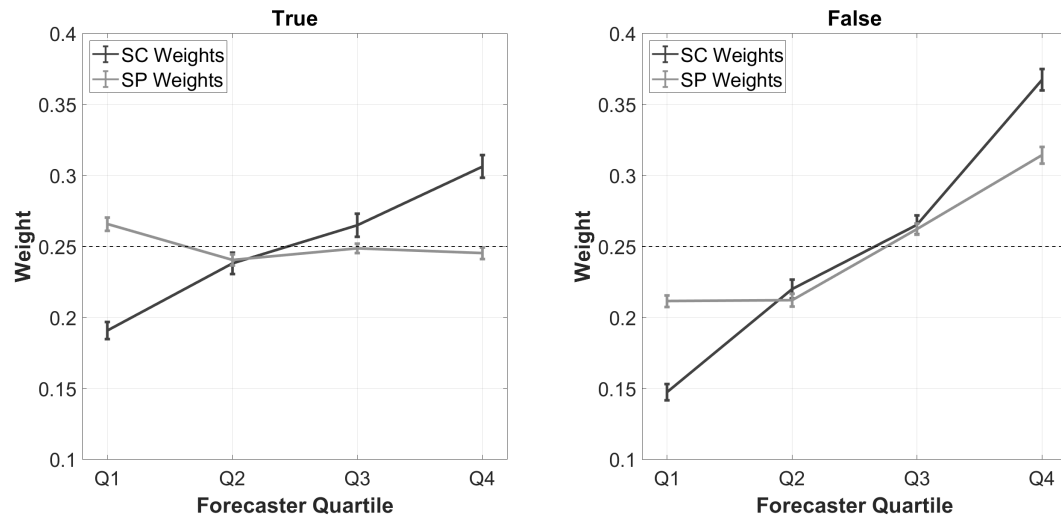


Figure 10 The average weight assigned by the SC algorithm and the SP algorithm as a function of forecasters' accuracy in the States dataset for the events where the outcome was "True" (left) and "False" (right). Forecasters were sorted into four bins, with 'Q1' containing the least accurate forecasters and 'Q4' containing the most accurate forecasters. The dotted line indicates the weights of an algorithm that weights all forecasters equally. Error bars represent the standard error.

We also applied the same analysis to the Quiz dataset from Experiment 2. We computed average weight assigned by the SC algorithm and SP algorithm assigned to each quartile of forecasters separately for each of the levels of difficulty and overall across all five levels of difficulty. As seen in Figure 11, both algorithms under-weight the least accurate forecasters and over-weight the most accurate forecasters in the more difficult datasets, but not necessarily in the easier dataset where a large proportion of forecasters are correct. On difficulty 2, the SC algorithm does a much better job at distinguishing between the best-performing and worst-performing individuals, whereas the SP algorithm assigns both groups approximately equal weights. Collapsing across all five difficulties (bottom right panel), the SC algorithm generates a larger aggregate weight for the most accurate forecasts and a smaller aggregate weight for the least accurate forecasters.

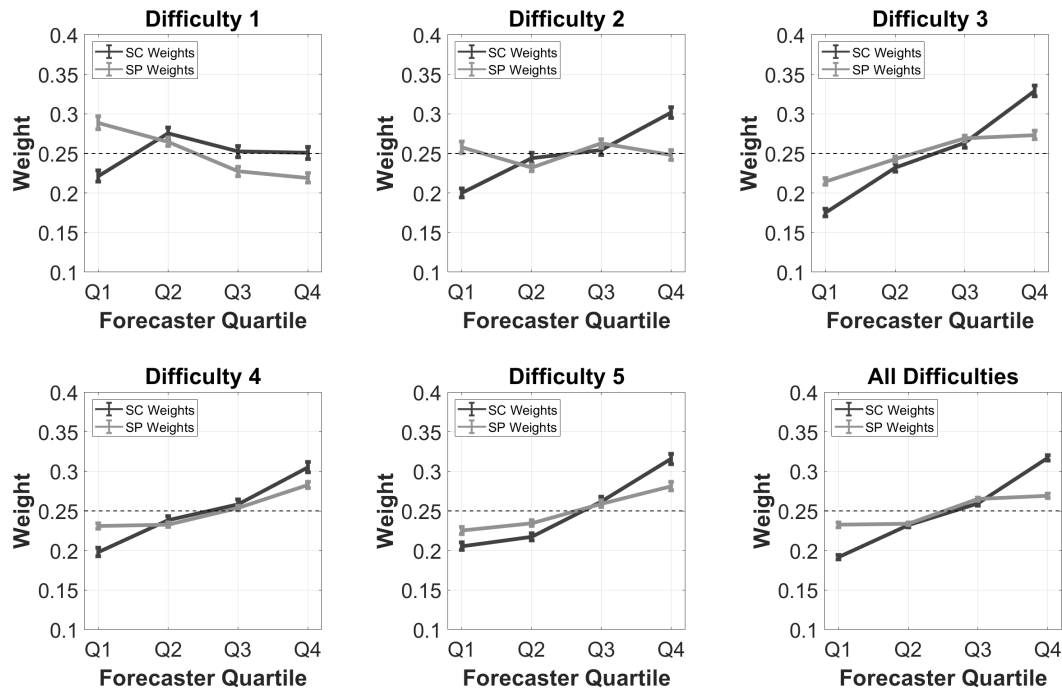


Figure 11 The average weight assigned by the SC algorithm and the SP algorithm as a function of forecasters' accuracy for each of the five individual difficulties and overall across all five difficulties in the Quiz dataset. Forecasters were sorted into four bins, with 'Q1' containing the least accurate forecasters and 'Q4' containing the most accurate forecasters. The dotted line indicates the weights of an algorithm that weights all forecasters equally. Error bars represent the standard error.

6.1.2 List of questions from Experiment 1: US States dataset from Chapter 2

Question	US state	Largest City	Capital City	Largest city is also the capital?
1	Alabama	Birmingham	Montgomery	False
2	Alaska	Anchorage	Juneau	False
3	Arizona	Phoenix	Phoenix	True
4	Arkansas	Little Rock	Little Rock	True
5	California	Los Angeles	Sacramento	False
6	Colorado	Denver	Denver	True
7	Connecticut	Bridgeport	Hartford	False
8	Delaware	Wilmington	Dover	False
9	Florida	Jacksonville	Tallahassee	False
10	Georgia	Atlanta	Atlanta	True
11	Hawaii	Honolulu	Honolulu	True
12	Idaho	Boise	Boise	True
13	Illinois	Chicago	Springfield	False
14	Indiana	Indianapolis	Indianapolis	True
15	Iowa	Des Moines	Des Moines	True
16	Kansas	Wichita	Topeka	False
17	Kentucky	Louisville	Frankfort	False
18	Louisiana	New Orleans	Baton Rouge	False
19	Maine	Portland	Augusta	False
20	Maryland	Baltimore	Annapolis	False
21	Massachusetts	Boston	Boston	True
22	Michigan	Detroit	Lansing	False
23	Minnesota	Minneapolis	Saint Paul	False
24	Mississippi	Jackson	Jackson	True
25	Missouri	Kansas City	Jefferson City	False
26	Montana	Billings	Helena	False
27	Nebraska	Omaha	Lincoln	False
28	Nevada	Las Vegas	Carson City	False
29	New Hampshire	Manchester	Concord	False
30	New Jersey	Newark	Trenton	False
31	New Mexico	Albuquerque	Santa Fe	False
32	New York	New York City	Albany	False
33	North Carolina	Charlotte	Raleigh	False
34	North Dakota	Fargo	Bismarck	False
35	Ohio	Columbus	Columbus	True
36	Oklahoma	Oklahoma City	Oklahoma City	True
37	Oregon	Portland	Salem	False
38	Pennsylvania	Philadelphia	Harrisburg	False
39	Rhode Island	Providence	Providence	True
40	South Carolina	Columbia	Columbia	True
41	South Dakota	Sioux Falls	Pierre	False
42	Tennessee	Memphis	Nashville	False
43	Texas	Houston	Austin	False
44	Utah	Salt Lake City	Salt Lake City	True
45	Vermont	Burlington	Montpelier	False
46	Virginia	Virginia Beach	Richmond	False

47	Washington	Seattle	Olympia	False
48	West Virginia	Charleston	Charleston	True
49	Wisconsin	Milwaukee	Madison	False
50	Wyoming	Cheyenne	Cheyenne	True

6.1.3 List of questions from Experiment 2: US Grades dataset from Chapter 2

Difficulty	Question	Outcome
5	In physics, work is measured in Joules	TRUE
2	Canine teeth are used for grinding food	FALSE
4	As the temperature increases, the solubility of gasses increases	FALSE
1	Magnets can stick to fridge doors because fridge doors contain a lot of plastic	FALSE
3	The deltoid muscle is located at the shoulders	TRUE
3	Amphibians are warm blooded	FALSE
2	The liver removes toxins from the blood	TRUE
1	The moon shines at night because it reflects light from the sun	TRUE
5	Microwaves contain more energy than visible light	FALSE
5	In physics, U-values measure how effective a material is an insulator	TRUE
3	Reptiles are vertebrates	TRUE
1	Rhinoceros are carnivores	FALSE
4	Asexual reproduction produces genetically identical offspring	TRUE
5	The acceleration of an object depends on the mass of the object	TRUE
4	Hurricanes usually only form if the sea temperature is less than 80 degrees Fahrenheit	FALSE
5	Heat transfers through outer space by radiation	TRUE
1	Rabbits are omnivores	FALSE
3	Butter contains little or no protein	TRUE
2	Materials that let electricity pass through them easily are called insulators	FALSE
5	In physics, kinetic particle theory links the movement of particles to chemical energy	FALSE
3	Mercury is the closest planet to the Sun	TRUE
2	Seasons are caused by the Earth's rotation around the Sun	FALSE
1	Ducklings are adult ducks	FALSE
4	White blood cells produce antigens to neutralize pathogens	FALSE
5	Kryptonite is commonly used as fuel in nuclear reactors to generate energy	FALSE
1	Crocodiles are mammals	FALSE
3	Carbon dioxide in the air is absorbed by plants	TRUE
3	The spleen filters blood and recycles old red blood cells	TRUE
2	Dandelions disperse their seeds by wind	TRUE
2	In physics, force is measured in Newtons	TRUE
4	Voluntary muscles are controlled by the cerebrum	TRUE
5	The rate at which an object transfers thermal energy depends on its surface area	TRUE
4	Hydrogen is a noble gas	FALSE
4	A country's carbon footprint measures their impact on the environment	TRUE
4	Epinephrine is a hormone prescribed to treat diabetes	FALSE
3	The most abundant gas in the Earth's atmosphere is Nitrogen	TRUE
5	Convex lenses are used to correct for short-sightedness	FALSE
4	The mass of an element equals the number of protons and electrons in one atom	FALSE
2	Sedimentary rocks are formed from cooled lava or magma	FALSE
1	Pitch describes the loudness of sounds	FALSE
3	The female reproductive cell is called a zygote	FALSE
3	Mitochondria are the cell's control centre	FALSE
5	Gamma radiation is the most dangerous type of nuclear radiation	TRUE
2	Baby horses are called ponies	FALSE

2	Respiration is the life process by which an organism synthesizes nutrients from sunlight	FALSE
3	Electrons are found at the centre of an atom	FALSE
5	Alpha radiation is commonly used to sterilize medical equipment	TRUE
1	A famine is when there is not enough food for people to eat	TRUE
1	The speed of an object equals the distance it is travelling over time	TRUE
1	Tadpoles lose their tails when they change into an adult frog	TRUE
1	Elephants are carnivores	FALSE
5	Heat transfer between objects occurs faster as the temperature difference between them decreases	FALSE
1	Water boils at 100 degrees Celsius at sea level	TRUE
5	When a charge flows through a resistor, it becomes positively charged	FALSE
2	Penguins are not naturally found in the north pole	TRUE
4	A catalyst has no effect on the equilibrium of a chemical system	TRUE
3	Red blood cells help your body fight pathogens	FALSE
4	In an ammonia molecule, hydrogen and nitrogen atoms share electrons	TRUE
3	"Co" is the symbol for Copper in the periodic table	FALSE
5	In physics, pressure is measured in Coulombs	FALSE
2	A solar eclipse occurs when the earth blocks the sun's light from reaching the moon	FALSE
5	Atoms of an element always have the same number of protons	TRUE
3	Joints attach muscles to the skeleton	FALSE
2	Herbivores are also known as primary consumers in the food chain	TRUE
1	Giraffes are born with long legs and necks	TRUE
5	In physics, electrical current is measured in Volts	FALSE
4	Electrons are lost or gained by atoms in ionic bonding	TRUE
1	A two year-old human is a toddler	TRUE
5	Infrared radiation has a longer wavelength than visible light	TRUE
4	In humans, the adrenal glands are found in the brain	FALSE
5	As the speed of a car increases, air resistance decreases	FALSE
1	Birds are not reptiles	FALSE
3	Animal cells do not contain chloroplasts	TRUE
3	Polar bears have a thick layer of fat for insulation	TRUE
1	Snails are animals	TRUE
2	Newts are commonly found in a desert environment	FALSE
2	Magnets are commonly made of iron	TRUE
3	Glass is an electrical conductor	FALSE
4	Earthquakes most commonly occur at the centre of tectonic plates	FALSE
1	Calves are baby cows	TRUE
2	Cellular respiration is the process of turning food into energy	TRUE
2	Cotton is a synthetic material	FALSE
1	Crabs have special front legs known as claws	TRUE
4	Antibiotics are ineffective against viruses	TRUE
1	Farm animals typically sleep on beds made of wool	FALSE
3	Light travels faster than sound	TRUE
5	The direction of a magnetic field is perpendicular to a wire carrying a current	TRUE
2	Ultraviolet light is invisible to the human eye	TRUE
5	The loudness of a sound wave is determined by its frequency	FALSE

3	Bleach is a strong acid	FALSE
2	A lack of Vitamin C causes scurvy	TRUE
4	During photosynthesis, chemical energy is converted into light energy	FALSE
4	A subsistence farmer is a farmer who focuses on raising animals and growing crops to feed their family	TRUE
4	Selective breeding results in an increased risk of diseases in the population	TRUE
4	Elements in the standard periodic table are arranged in terms of atomic mass	FALSE
1	Baby chickens are called hens	FALSE
2	The retina is the nerve that takes signals from the eye to the brain	FALSE
4	Combustion reactions are typically irreversible	TRUE
2	Snakes need to shed their skin regularly as they grow	TRUE
3	Penicillin, yeast, and mold are examples of bacteria	FALSE
2	Gestation is another name for pregnancy	TRUE
4	Darwin's theory was not widely accepted when it was first published in the late 19th century	TRUE
2	Sandpaper is typically made of sheets of paper or cloth coated with abrasive material	TRUE
3	"Cl" is the symbol for chlorine in the periodic table	TRUE
2	It takes the Earth one year to rotate once on its axis	FALSE
1	Flashlights produce light from electricity stored in batteries	TRUE
5	A cell or a battery produces an alternating current	FALSE
4	GNP measures total value of goods and services produced within the country in a single year	FALSE
5	Infrared radiation is typically used to sterilize medical equipment	FALSE
1	Hamsters hate to run	FALSE
1	The elbow is a joint	TRUE
1	Baby eagles typically learn to fly before they leave the nest	TRUE
2	An impermeable material is one that allows water to pass through it easily	FALSE
2	Rice is typically a good source of carbohydrates	TRUE
2	In winter, nights are longer than days	TRUE
2	Electromagnets are temporary magnets	TRUE
4	The amount of water in the body is regulated primarily via thyroid gland	FALSE
4	Hurricanes usually form over tropical seas	TRUE
2	Flour is made from ground grains or roots	TRUE
3	The heart is made of muscle cells known as smooth muscle	FALSE
1	Lizards typically have six legs	FALSE
5	The center of mass of a regular-shaped object is always located in the center of the object	FALSE
2	The sides of a river are called the channels	FALSE
5	Thermal energy naturally only flows from colder to hotter objects	FALSE
1	Magnets attract plastic	FALSE
1	Caterpillars turn into pupa before becoming a butterfly	TRUE
4	Fractional distillation uses the tensile strength of substances in crude oil to separate them	FALSE
4	Hydrogen gas is typically given off when group 1 elements react with water	TRUE
5	In physics, the weight of an object is equal to the object's mass times gravity	TRUE
4	An alkali is a substance that produces hydrogen ions when added to water	FALSE

5	Ultrasound refers to sound waves with frequencies below the audible limit of human hearing	FALSE
5	Convection of heat cannot occur for solids	TRUE
3	Pupils constrict depending on the amount of light	TRUE
1	Soap bubbles are mostly filled with water	FALSE
2	Nylon is a synthetic material	TRUE
2	When water boils it turns into steam	TRUE
2	Water is an opaque material	FALSE
1	Clay softens when heated in a kiln	FALSE
3	A full moon occurs when the Sun is completely illuminated by the Moon	FALSE
3	The muscular wall that separates the left and right sides of the heart is called the ventricle	FALSE
4	Increasing the number of impermeable rocks in rivers help decrease the flood risk	FALSE
1	Tortoises do not have flippers	TRUE
5	Fiber optic broadband is an example of an application of total internal refraction	TRUE
1	Crows build their nests in trees	TRUE
1	Children can typically walk by the time they become a toddler	TRUE
1	Baby kangaroos are called pups	FALSE
5	In physics, force equals mass times velocity	FALSE
5	Step-down transformers help reduce the voltage supplied to each appliance from the mains electricity	TRUE
2	Infrared light is invisible to the human eye	TRUE
3	Ribosomes are small organelles composed of RNA	TRUE
5	A vacuum can be used to induce a potential difference at the ends of a coil of wire	FALSE
2	Rubber is not attracted to magnets	TRUE
3	The top layer of the earth is called the crust	TRUE
4	Brass is an alloy containing primarily copper and zinc	TRUE
3	The female reproductive cell is called a blastocyst	FALSE
4	The maximum number of electrons in the first electron shell is 2	TRUE
1	It is possible to smell scents inhaled only through the mouth	TRUE
5	Energy is measured in Joules	TRUE
5	Infrared radiation generally cannot pass through walls	TRUE
3	Strong alkalis have a sour taste	FALSE
3	Incisors are the eight teeth in the front and center of your mouth	TRUE
2	Amperes are a measure of mass	FALSE
3	The muscle system makes up about half the weight of the average male body	TRUE
4	Food cooks faster in boiling vegetable oil than in boiling water	TRUE
4	One disadvantage of biofuels is that they are carbon neutral	FALSE
4	Height is an example of discontinuous variation in a trait	FALSE
4	Hormones are transported around the body via the peripheral nervous system	FALSE
3	The heart is divided into multiple chambers	TRUE
2	Oxygen is a product of respiration	FALSE
2	Crocodiles give birth to live young	FALSE
2	Metres is a measure of volume	FALSE
3	Argon is a halogen	FALSE
5	In physics, Coulombs are the units for power	FALSE
3	The smallest bone in the human body is found inside the wrist	FALSE

1	Hot steel is easier to shape than cold steel	TRUE
4	Pressure directly affects the solubility of gases in water	TRUE
3	A pH number of 14 indicates a strong base	TRUE
5	In physics, kinetic energy is equal to mass times velocity	FALSE
1	Magnets commonly have three poles	FALSE
2	Glass is a metal	FALSE
5	Resistance in a circuit can be calculated by measuring current and voltage	TRUE
5	In physics, the unit for charge is Ohms	FALSE
1	Gorillas typically live in forests	TRUE
4	Voluntary muscle messages are processed in the medulla of the brain	FALSE
5	Heat exchangers can be used to convert wasted heat energy into a useful source of power	TRUE
3	Rock formations found in caves that extend upwards are called stalactites	FALSE
4	When contours on a map are very close together, it indicates a steep slope	TRUE
2	Vitamin B can be synthesised by the body from exposure to sunlight	FALSE
3	Caveology is the science that involves studying caves and exploring them	FALSE
5	Wood is a poor conductor of heat	TRUE
4	The hormone progesterone is secreted by the ovaries	TRUE
1	Insects are part of Rabbits' diets	FALSE
1	Frogs usually lay their eggs in sand	FALSE
3	"Ca" is the symbol for calcium in the periodic table	TRUE
3	The testes produce a hormone called testosterone	TRUE
1	Herbivores eat both plants and animals	FALSE
5	An advantage of overhead electrical lines compared to underground lines is that they are cheaper to repair	TRUE
4	Aluminum is a widely used metal for its light and malleable properties	TRUE
3	Babies are considered full term at 30 weeks	FALSE
4	Decane splitting into octane and ethene is an example of a polymerization reaction	FALSE
1	Fish are not considered animals	FALSE
2	The source of power in an electrical circuit comes from the battery	TRUE
3	A newly fertilized human egg cell is called a zygote	TRUE
2	Oak trees produce acorns	TRUE
4	Skin and bones are two organs that form part of the auxiliary nervous system	FALSE
4	Penicillin was the first antibiotic to be discovered	TRUE
4	Ionic bonding holds the atoms together in water	FALSE
5	An increase in current through a wire exposed to a magnetic field will also increase the force experienced by the wire	TRUE
1	Goslings are baby ducks	FALSE
5	As a substance changes state from liquid to gas, the amount of energy particles have increases	TRUE
2	The Moon's orbit of the Earth lasts approximately 24 hours	FALSE
4	Suffocation is the main cause of death from avalanches	TRUE
3	Hedgehogs are nocturnal and hibernate during the winter	TRUE
5	In physics, the momentum of an object is equal to the mass of the object multiplied by its velocity	TRUE
1	Ferns typically grow on the ground in forests	TRUE
5	If a person's mass is 65 kilograms, their weight on Earth would be close to 65 newtons	FALSE

2	Gold, iron, and copper are all types of metals	TRUE
1	The speed of an object is measured in Volts	FALSE
2	The spinal nerves take signals from the ear to the brain	TRUE
4	Covalent bonding holds the atoms together in sodium chloride	FALSE
3	Sound frequency is measured in Joules	FALSE
4	Sensory neurons are responsible for sending messages from the brain to different glands in the body	FALSE
5	When a vehicle is in motion, the air resistance it experiences is a much lower force than the friction with the road	FALSE
2	Flowering plants produce many seeds which never grow	TRUE
4	Photosynthesis is an example of an endothermic reaction	TRUE
2	Magnets are usually made of plastic	FALSE
2	Mass is measured in Newtons	FALSE
4	The addictive substance in alcohol is called propanol	FALSE
2	Butter is liquid at room temperature	FALSE
2	Balloons are examples of gases trapped inside a solid	TRUE
5	In physics, pressure is equal to unit area divided by force	FALSE
4	Scurvy and anemia are diseases not caused by bacteria or viruses	TRUE
5	In a physics, current is defined as the flow of electrons through a circuit	TRUE
3	Plasma cells carry and release antibodies to help fight foreign pathogens	FALSE
3	Lithium is the only metal that is liquid at room temperature	FALSE
5	In physics, energy can be transferred, stored, and dissipated	TRUE
1	The leaves of the carrot plant are the most common part of the carrot plant eaten by humans	FALSE
2	Light travels at approximately 300,000 miles per second in a vacuum	FALSE
5	In a circuit, power refers to the rate at which energy is absorbed	FALSE
4	The estimated age of the earth is approximately 4.5 billion years	TRUE
2	Sound waves travel down the ear canal before hitting retina	FALSE
2	The moon generates its own light	FALSE
3	Light travels slower in air than water	FALSE
4	Earthquakes and volcanoes typically occur at the boundaries of tectonic plates	TRUE
5	If a charge flows through a resistor in a circuit, the resistor becomes negatively charged	FALSE
1	Baby swans are called ducklings	FALSE
1	Squirrels commonly live in trees	TRUE
3	When a female's body releases an egg it is known as ovulation	TRUE
4	Organisms working together to gain an advantage in competition is an example of parasitism	FALSE
5	In physics, the rate of evaporation is determined by both temperature and the surface area of the liquid	TRUE
3	The hip is an example of a hinge joint	FALSE
1	Trampolines typically contain dozens of coiled springs	TRUE
1	Baby leopards are called calves	FALSE
1	Most wood is typically harder than steel	FALSE
2	In plants, germination is when seeds leave the parent plant	FALSE
1	Alligators are mammals	FALSE
1	Walls are considered opaque because they do not let light through	TRUE

3	Lymphocytes are a type of white blood cell that help your body fight pathogens	TRUE
3	"Fe" is the symbol for iron in the periodic table	TRUE
4	Nitrogen can typically form up to two covalent bonds	FALSE
2	Mitosis and Meiosis are two types of cell division	TRUE
3	Our solar system is located in the Milky Way galaxy	TRUE
2	Cotton is a material harvested from animals	FALSE
2	Steel is made from a mixture of iron and other elements	TRUE
5	The frequency of a simple pendulum that oscillates 5 times per second is 5,000 Hertz	FALSE
3	Both penguins and sharks have developed gills for breathing underwater	FALSE
1	Rowdy means the opposite of noisy	FALSE
5	The terminal velocity of an object is reached when the object's acceleration exceeds 0	FALSE
1	Butterflies usually lay their eggs on leaves	TRUE
4	Hydrochloric acid in the stomach is an example of passive immunity in the body	TRUE
1	There are receptors for temperature on most of our skin	TRUE
4	The atomic number of an element tells us the number of neutrons in the nucleus	FALSE
3	The kneecap is also known as the radius	FALSE
4	Smallpox was the first disease for which a vaccine was produced	TRUE
5	Matte, black surfaces are typically the best reflectors of heat	FALSE
5	In physics, U-Values measure how effective a material is at absorbing carbon dioxide	FALSE
3	"Br" is the symbol for bronze in the periodic table	FALSE
3	Food from the esophagus enters the stomach	TRUE
1	Woodpeckers live in trees	TRUE
1	Crabs are amphibians	FALSE
2	Marble is a naturally-occurring material	TRUE
1	Water freezes at 0 degrees Celsius at sea level	TRUE
2	Seals, sharks, and whales are all classified as mammals	FALSE
3	Adaptation is when species develop characteristics over time which allow them to survive better in their environment	TRUE
4	The vacuoles of the cells are the main organelle involved when cloning animals from adult cells	FALSE
3	Plasma is the fluid in the body which is a mixture of water, minerals, nutrients, proteins, and other substances	TRUE
3	Canine teeth are used to hold and tear apart food	TRUE
5	X-rays are part of the electromagnetic spectrum	TRUE
3	The kidney produces acids and enzymes to help to kill bacteria and other infectious organisms consumed by a person	FALSE
5	The force experienced by a current-carrying wire can be reversed by reversing the direction of current/magnetic field	TRUE
4	The last ice age occurred during the Jurassic period	TRUE
5	Porcelain does not allow electricity to flow through it easily	TRUE
2	Mosses are classified as plants	TRUE
1	Lights contain a bulb that glows to produce light	TRUE
4	Electrons are found in the nucleus of a atom	FALSE
1	Matchboxes contain flammable material	TRUE
4	The main impact of deforestation is the loss of habitat for various living species	TRUE
3	White blood cells are also known as cytoplasm	FALSE

5	If an elastic material passes its point of proportionality, it will lose its elasticity and not return to its original size and shape	TRUE
5	In physics, the current of a resistor is inversely proportional to the potential difference across it	FALSE
2	Heavier objects have less momentum than lighter objects moving at the same speed	FALSE
1	Foxes have wool to keep them warm	FALSE
3	"Mg" is the symbol for magnesium in the periodic table	TRUE
5	Sound waves and electromagnetic waves are examples of longitudinal waves	FALSE
1	Convex lenses are commonly used to help people with hearing disabilities	FALSE
4	Hydrocarbates are essential to a balanced human diet	FALSE
2	Jaguars are commonly found on the American continents	TRUE
4	Energy changes in chemical reactions are commonly measured in kilowatts	FALSE
1	Baby tigers are called kittens	FALSE
1	Rubber is commonly used for its elastic properties	TRUE
3	The largest lymphatic organ is the liver	FALSE
2	Pandas originate from Egypt	FALSE
4	A substance with a pH of 8 is a strong acid	FALSE
3	In physics, pressure is measured in Pascals	TRUE
3	Stored energy due to gravity is called potential energy	TRUE
3	The nucleus is the cell's control centre	TRUE
5	When two different insulating materials are rubbed against each other, they become electrically charged	TRUE
3	Sandstone (which is grainy, crumbly and may contain fossils) is an example of sedimentary rock	TRUE
4	Buildup of glucosic acid is responsible for causing cramps after exercise	FALSE
2	Light travels more slowly through air than glass	FALSE
4	Selective breeding results in a smaller gene pool	TRUE
2	Insects and wind are common methods of pollination in plants	TRUE
3	Antibodies are the immune system's first line of defense	FALSE
5	In a circuit, a fuse can be reset after it is triggered	FALSE
3	The central nervous system consists of the brain and the spinal cord	TRUE
1	Batteries store chemical energy	TRUE
4	Increased pressure speeds up the rate of chemical reactions in both gases and liquids	FALSE
2	In physics, gravity is a force	TRUE
5	In physics, the stability of an object can be increased by lowering its center of mass	TRUE
2	Woodlice typically live in dark places	TRUE
2	The ovary of a plant grows into a flower	FALSE
3	The loudness of sound is measured in decibels	TRUE
1	At night, you cannot see the Sun	TRUE
2	Centimetres are a measure of length	TRUE
3	Arteries also carry lymphatic fluid to the capillaries	FALSE
4	Tsunamis are common secondary effects from earthquakes	TRUE
1	Dolphins typically live in swamps	FALSE
5	Coal, gas, and oil are all examples of fossil fuels	TRUE
1	Sounds that are too loud can damage our hearing	TRUE
5	Xenon is a fuel commonly used in nuclear power	FALSE
1	Bats are classified as birds	FALSE

4	There are four covalent bonds involved in a methane molecule	TRUE
4	Sedimentary rock typically form as a result of extremely high pressure or heat	FALSE
5	In physics, the turning effect of a force is known as a moment	TRUE
4	Nitrogen is the element that has 6 electrons	FALSE
5	Knowing an appliance's power consumption and potential difference would allow someone to calculate the current	TRUE
5	The purpose of the turbine in a power station is to turn the generator which produces electricity	TRUE
1	Balloons containing helium will float higher in air than balloons containing carbon dioxide	TRUE
2	Wool is commonly used for knitting	TRUE
4	Bacteria are the primary source of energy for most food chains	FALSE
1	The knee is a muscle	FALSE
4	Low atmospheric pressure lead to periods of drought	FALSE
2	Cactuses have modified leaves known as spines	TRUE
4	Carbon dioxide is an essential component required for photosynthesis	TRUE
1	Planes have wheels that help them move on the ground	TRUE
5	The smaller the heat difference between two objects, the faster the heat transfer between them	FALSE
1	Modern trains usually run on steam	FALSE
2	Kangaroos are commonly found in South Africa	FALSE
4	Some vaccines contain the live pathogen that has been treated to make it harmless	TRUE
2	School playgrounds are commonly made of tarmac or concrete	TRUE
2	Iron, copper, and nylon are all examples of metals	FALSE
4	An increase in pH is an example of an abiotic change in the environment	TRUE
4	The surface area to volume ratio for nanoparticles is much greater than that of larger particles	TRUE
3	Red blood cells are also known as leukocytes	FALSE
2	Pollination describes the moment a seed begins to grow	FALSE
1	Birds commonly eat seeds from plants	TRUE
4	Most of the Nitrogen that plants use are absorbed through their roots	TRUE
2	Plants make their own food from sunlight	TRUE
1	The engine of a car is typically found in the back of the car	FALSE
2	Erosion occurs when rivers deposit rocks and soils as they flow	FALSE
1	Cats often lick their own fur	TRUE
5	Geothermal energy refers to heat energy produced in underground nuclear power stations	FALSE
4	The first two electron shells in Neon are fully filled with electrons	TRUE
3	White blood cells form blood clots when bleeding occurs	FALSE
5	Kryptonite is a fuel commonly used in nuclear reactors	FALSE
3	The cerebrum makes up the majority of the brain	TRUE
3	The lungs are the largest organ of the body	FALSE
3	Junctions in the nervous system are known as axons	FALSE
4	When international aid is given directly from one country to another, it is called multilateral aid	FALSE
1	Cakes are commonly made from flour, butter, and eggs	TRUE

5	If the voltage in a circuit remains constant but the resistance is increased, current decreases	TRUE
1	Fish have fur to keep them warm	FALSE
5	Refraction can be defined as the change in direction of light as it passes from one medium to another	TRUE
5	In physics, DC electricity refers to a current that can travel bi-directionally	FALSE
3	Bile is produced in the gall bladder and stored in the liver	FALSE
5	Infrared radiation travels at a speed slower than wavelengths of light with higher frequencies	FALSE
2	Pollen grains are typically stored underneath a plant's petals	FALSE
5	For an elastic material, Boyle's law describes the relationship between stretch and the force applied	FALSE
2	The south pole of a magnet will repel the north pole of another magnet	FALSE
1	Flour is made from wheat	TRUE
5	In physics, work done is equal to the force needed to move an object multiplied by the distance it moved	TRUE
5	Aluminum is typically a good insulator against heat loss	FALSE
3	The hip bone is an example of the axial skeleton	FALSE
5	A sound wave with a long wavelength will have a low pitch	TRUE
3	The majority of elements in the periodic table are metals	TRUE
4	The core of the earth is the hotter than its outer layers	TRUE
3	The human male sex cell is known as the gonads	FALSE
2	The opposite of rough is smooth	TRUE
1	Plastics are generally less flexible than wood	FALSE
3	Cooled magma is not a common component of limestone	TRUE
2	Modern roads are commonly covered in tarmac	TRUE
2	Opaque means the opposite of transparent	TRUE
3	The diaphragm is located underneath the lungs and functions to help us breathe	TRUE
5	In physics, the unit for resistance is Ohms	TRUE
4	The kiwi is an extinct flightless bird that lived on the island of Mauritius	FALSE
1	Candles are commonly made of wax	TRUE
2	Litres is a measure of length	FALSE
5	In kinetic particle theory, all collisions within a system are assumed to be elastic	TRUE
1	Our tongues have several types of taste receptors	TRUE
5	Infrared radiation was first discovered over 200 years ago	TRUE
4	Secondary industries dominate the market in emerging economies	FALSE
5	The two most common forms of waste energy are heat and kinetic energy	FALSE
2	The seeds of a flowering plant grow into fruit	FALSE
5	The Doppler effect describes a decrease in the frequency of a wave as the source and observer meet	FALSE
1	Piano keys increase in pitch when moving from left to right	TRUE
1	Yachts move primarily due to the ocean currents	FALSE
2	Mosses are rarely found in desert environments	TRUE
2	Rubber is a poor conductor of electricity	TRUE
5	Shiny, white surfaces are typically poor absorbers of heat	TRUE
2	Plastic is a good conductor of electricity	FALSE
3	Molars (teeth) are mostly used to hold and tear food apart	FALSE

2	Respiration is the life process by which food is transformed into energy	FALSE
2	A brittle material is one that absorbs liquids easily	FALSE
3	"Na" is the symbol for Sodium in the periodic table	TRUE
3	Blood exits the human heart through the ventricles	TRUE
5	Electrostatic force is the centripetal force responsible for the circular motion of planets orbiting the Sun	FALSE
3	The cell membrane separates the interior of a cell and the exterior fluid surrounding a cell	TRUE
3	The voice box is also known as the epiglottis	FALSE
2	Carbohydrates provide the body with a source of energy	TRUE
3	The Cochlea is part of the outer ear	FALSE
3	Phosphorus is a metal	FALSE
4	Plant cells are easier to clone than animal cells	TRUE
3	The anvil is an inner ear structure filled with fluid	FALSE
3	The Arctic Hare change their coat colour depending on the season	TRUE
3	Neutral substances have a pH of 1	FALSE
2	Ceramics are not considered metals	TRUE
1	Metals are bad at absorbing heat	FALSE
1	Fish commonly have multiple types of fins	TRUE
3	Images appear on the retina upside down	TRUE
4	It is common for the dominant male lion in prides to chase some of the male cubs away as they approach sexual maturity	TRUE
4	Estrogen is commonly produced in the adrenal glands	TRUE
2	Caterpillars become a cocoon before turning into butterflies	TRUE
1	Daffodils are usually blue in color	FALSE
5	In physics, evaporation can be described as low energy particles breaking free of the liquid at the surface	FALSE
2	The Earth's orbit of the sun lasts approximately 24 hours	FALSE
3	When the triceps contract, they extend the leg at the knee joint	FALSE
4	Isotopes have the same number of protons, but different number of neutrons	TRUE
4	Asexual reproduction produces genetically identical offspring due to fusion of gametes	FALSE
2	In chemistry, a solution is when one material dissolves in another	TRUE
1	Objects that contain a lot of air float easily in water	TRUE
1	Most plants are red in color	FALSE
2	Reindeer are native to northern Europe	TRUE
2	Green objects absorb green light, but reflect all other colours	FALSE
3	The pitch of a string increases as the vibration frequency increases	TRUE
5	Electrical current is defined as the flow of electric charge	TRUE
3	The most reactive group 1 element in the periodic table is Lithium	FALSE
5	The amount of electrical energy transferred to an appliance depends on its power and the length of time it is switched on	TRUE
4	Subsistence farming is when crops are cultivated by farmers to be sold commercially	FALSE
3	Fingers and toes have joints called ball-and-socket joints	FALSE
1	Eagle owls are classified as birds	TRUE
1	Hearing aids are commonly used by people with hearing disabilities	TRUE
5	Both rough and smooth surfaces experience the same level of frictional force	FALSE

1	The velocity of an object is measured in hours per minute	FALSE
2	Glass is usually malleable	FALSE
3	The pulmonary artery pumps blood back into the left ventricle of the heart	FALSE
5	The size of an electric current depends on the flow of neutrons in the circuit	FALSE
4	Organisms that are tolerant of high levels of salt, high temperatures, or high pressures are called extremophages	FALSE
4	Filtration is commonly used to accurately measure volumes of liquids required for chemical reactions	FALSE
4	Antibiotic resistance in bacteria arises due to random mutations	TRUE
2	Beans are a good source of protein	TRUE
1	Most wavelengths of visible light cannot pass through transparent objects	FALSE
5	Watts refers to the flow of Joules per second	TRUE
5	Specific heat capacity refers to the amount of heat needed to raise a system's temperature by one degree	TRUE
1	Omnivores only eat meat	FALSE
4	Electrolysis is the process of splitting substances up using heat energy	FALSE
4	Random mutations in DNA are a common cause of mass extinction in a species	FALSE
1	Spiders typically have twelve legs	FALSE
3	The brain's frontal lobe is responsible for most higher-order functions such as planning, problem-solving, judgement, and creative thought.	TRUE
4	All organic compounds contain the element carbon	TRUE
5	If the voltage in a circuit is increased whilst the resistance remains constant, the current stays the same	FALSE
4	Enzymes are commonly used to cut out specific genes to place into other organisms	TRUE
1	Children typically learn to talk before they can walk	FALSE
1	Glass is commonly used in construction because it is opaque	FALSE
4	Deforestation is considered one of the main contributors to the greenhouse effect	TRUE
5	A kettle will have a power of 100 watts if it transfers 1,000 joules of energy in 10 seconds	TRUE
2	Aluminum is made from melted plastics	FALSE
4	Eye color is an example of continuous variation in a trait	FALSE
4	Transition elements are commonly used as catalysts in chemical reactions	TRUE
1	Tadpoles hatch from frogs' eggs	TRUE
4	Metallic bonds are typically stronger than covalent bonds	FALSE
1	Tomatoes grow on plants	TRUE
5	In physics, the unit for current is Volts	FALSE
5	As you dive deeper into the ocean, the hydrostatic pressure of the water increases	TRUE
5	The force applied to a current-carrying wire can be increased by decreasing the magnetic field	FALSE
1	Daylight is commonly used for drying clothes	TRUE
3	Epilepsy is a disorder of the nervous system	TRUE
2	Condensation is when a solid turns into a gas	FALSE
4	The pituitary gland is found in the brain	TRUE
5	A concave lens magnifies the perceived size of an object	FALSE
3	In the periodic table, each period refers to a horizontal row	TRUE

6.2 Chapter 3 Appendices

6.2.1 Manuscript adapted from Chapter 3

This manuscript was largely adapted from the content in Chapter 3.

1

2

3 Using meta-predictions to identify experts in the crowd when past

4 performance is unknown

5

6

7 Marcellin Martinie,^{1*} Tom Wilkenin², Piers D. L. Howe,¹

8

9 ¹ Melbourne School of Psychological Sciences, The University of Melbourne, Parkville 3010,

10 Victoria, Australia.

11 ² Department of Economics, The University of Melbourne, Parkville 3010, Victoria, Australia.

12

13 *Corresponding author.

14 E-mail: marcellin.martinie@unimelb.edu.au (MM)

15

16

17 Abstract

18 A common approach to improving probabilistic forecasts is to identify and leverage the
19 forecasts from experts in the crowd based on forecasters’ performance on prior questions with
20 known outcomes. However, such information is often unavailable to decision-makers on many
21 forecasting problems, and thus it can be difficult to identify and leverage expertise. In the current
22 paper, we propose a novel algorithm for aggregating probabilistic forecasts using forecasters’
23 meta-predictions about what other forecasters will predict. We test the performance of an
24 extremised version of our algorithm against current forecasting approaches in the literature and
25 show that our algorithm significantly outperforms all other approaches on a large collection of
26 500 binary decision problems varying in five levels of difficulty. The success of our algorithm
27 demonstrates the potential of using meta-predictions to leverage latent expertise in environments
28 where forecasters’ expertise cannot otherwise be easily identified.

29 1. Introduction

30 The fact that judgments can be improved by aggregating predictions across forecasters in a
31 crowd has been well-known for over a century [1]. Simple averaging is a common approach to
32 aggregating probabilistic forecasts and works well when forecasters have the same level of
33 expertise. However, in practice, expertise is rarely constant across forecasters [2, 3]. A number of
34 aggregation approaches have been developed to identify and leverage differences in expertise
35 using forecasters’ past performance on questions with known outcomes [4, 5] and forecasters’
36 past contributions to the crowd forecast [6]. Unfortunately, information regarding past
37 performance may often be unavailable because collecting forecasters’ responses to a set of
38 relevant questions can be very time-consuming, costly, or otherwise impractical.

39 In a recent paper, Prelec, Seung, and McCoy [7] developed an innovative algorithm that
40 uses meta-predictions—predictions about what others will predict—to correct for biases in the

crowd where information regarding past performance is unknown. Their surprisingly popular (SP) algorithm predicts that the outcome that is more popular than the crowd expects (i.e., the surprisingly popular outcome) to be the correct answer.

In the current paper, we explore an alternative way of using meta-predictions to improve probabilistic forecasts. We propose the meta-probability weighting (MPW) algorithm, which weights the probabilistic forecasts of each forecaster by using the absolute difference between their prediction and their meta-prediction of the average prediction of others. As shown in our theoretical model discussed in the S1 Appendix, the weight assigned to each forecaster in the MPW algorithm is proportional to the absolute difference between the forecaster’s prior and the forecaster’s posterior in a Bayesian framework where forecasters receive private signals and share a common prior. Thus, forecasters with more informative private signals will be weighted more in the algorithm than those with less informative signals. Although this reweighting does not guarantee that the probabilistic forecast generated by the meta-probability weighting algorithm is closer to the truth than the simple average on a question-by-question basis, it does ensure that experts—individuals who have access to a more informative information system—will have higher expected weights than novice in crowds containing both types of individuals. Since experts will have better forecasts than novices on average, we hypothesize that the MPW algorithm will yield better probabilistic forecasts in the aggregate across many problems.

We test the performance of an extremised version of our algorithm against three current forecasting approaches in the literature—the extremised simple average, an extremised version of the minimal pivoting procedure of Palley and Soll [13], and the p''_{cs} aggregator of Satopää, Pemantle, and Ungar [8]—using a large collection of 500 binary decision problems varying in five levels of difficulty. As discussed below, these alternative algorithms aim to improve the aggregate forecasts by correcting for the sharing or overlap in common information between forecasters. We find that the new algorithm outperforms all three alternative algorithms. We find

that this outperformance is driven by improved performance on more difficult questions where there is likely to be heterogeneity in expertise.

The rest of this paper is organized as follows. In Section 2, we provide a formal definition of the MPW algorithm and discuss the theoretical properties of the algorithm. In Section 3, we describe our experimental design, the analyses we plan to conduct, and formally define each alternative aggregation approach. In Section 4, we examine the performance of each aggregation approach both generally and at the dataset level. Finally, in Section 5, we review the implications of these findings and the contribution it provides to the literature. The S1 Appendix contains our theoretical model while the S2 Appendix contains a comparison of the MPW algorithm and alternatives using the NCAA Men’s basketball dataset of Palley and Soll [13].

2. The MPW algorithm

Let X be the event space with events $X_1, X_2, \dots, X_K$ where K is the total number of events. Let $P_{i,k}$ be the probability forecast of the i^{th} forecaster for the k^{th} event and let $M_{i,k}^P$ be this forecaster’s meta-prediction of the average forecast of others. Then, the probabilistic forecast made by the MPW algorithm, $T_{MPW}(X_k)$, is given by

$$T_{MPW}(X_k) = \sum_{i=1}^{N_k} W_{i,k} P_{i,k} \quad (1)$$

where N_k is the total number of forecasters for the k^{th} event and

$$W_{i,k} = \frac{|P_{i,k} - M_{i,k}^P|}{\sum_{i=1}^{N_k} |P_{i,k} - M_{i,k}^P|}. \quad (2)$$

Note that by construction, the weights for each event k sum up to 1.

The weights for the MPW algorithm are informed by our theoretical model developed in the S1 Appendix. In our theoretical model, individuals share a common prior about the likelihood

85 that the answer is true and receive private signals from one of two information systems that are
86 ranked in terms of their informativeness. We allow the prior to be biased—as might be the case if
87 forecasters receive a commonly observed public signal and update their beliefs to an informed
88 common prior before receiving their private signals—but assume that signals are independent
89 after conditioning on the state. We also assume all forecasters have common knowledge about the
90 likelihood of a randomly selected forecaster receiving each potential signal in the true state and
91 the false state. This assumption implies that two forecasters who receive the same private signal
92 will have the same meta-prediction about the reports made by others.

93 We define an expert as an individual who receives a signal from the more informative
94 information system and a novice as an individual who receives a signal from the less informative
95 one. We show that under our theoretical assumptions, the weight of an individual is zero if the
96 individual's prior is equal to his or her posterior and that individual weights are increasing
97 linearly in the distance between a forecaster's prior and his or her posterior. In this sense,
98 individuals with more informative private signals will be weighted more than individuals with
99 less informative private signals. Since experts have a more informative signal than a novice on
100 average, we can use Blackwell's Theorem [21-25] to show that the expected weight of an expert
101 is greater than the expected weight of a novice. We predict that the overweighting of experts in
102 the algorithm will improve probabilistic forecasts in the aggregate.

103

104 **3. The Experiment**

105 To test the MPW algorithm, we conducted an online experiment where we presented
106 participants with US grade school true-or-false general science statements varying on five
107 predefined levels of difficulty. We selected problems which varied systematically in difficulty
108 because they provide a natural environment in which the level of expertise in the crowd varies

109 accordingly. Our theoretical model predicts that the MPW algorithm is likely to offer the greatest
110 improvement over simple averaging on moderate-difficulty and high-difficulty forecasting
111 problems, where crowds are likely to contain forecasters with latent expertise. In contrast, the
112 MPW algorithm is likely to provide little-to-no benefit over simple averaging on low-difficulty
113 problems, where most forecasters are likely to be experts.

114 **3.1 Experimental Design**

115 We generated 500 science statements at a US primary and secondary grade school level.
116 Questions and content were adapted from worksheets on the Education Quizzes website
117 (<http://www.educationquizzes.com/us>), and then converted into true or false statements.
118 Approximately 2-3 questions were taken from each worksheet from the Biology, Chemistry,
119 Geography, Physics, and General Science categories, spanning from grades 1 to 12, broken up
120 into five levels of difficulty (grades 1 and 2; grades 3, 4, and 5; grades 6, 7, and 8; grades 9 and
121 10; and grades 11 and 12). We coded “difficulty 1” as the lowest difficulty level, and “difficulty
122 5” as the highest difficulty level. We treated each set of 100 questions of the same difficulty as an
123 individual dataset. An example of a statement in difficulty 1 was “Omnivores only eat meat”. In
124 contrast, difficulty 5 contained statements such as “Sound waves and electromagnetic waves are
125 examples of longitudinal waves”. The full set of experiment questions, participant responses, and
126 analysis code (for the MATLAB program, please see
127 <https://www.mathworks.com/products/matlab.html>) are included in the supplementary
128 information files.

129 The experiment was approved by the Melbourne School of Psychological Sciences Human
130 Ethics Advisory Group (Ethics ID: 1647855.1) and all experiments were performed in accordance
131 with the relevant guidelines and regulations. We recruited 500 respondents from Amazon
132 Mechanical Turk; only respondents inside the US were able to participate in the experiment.

133 Participants were paid a flat fee of USD \$4.00 for completing the survey and all participants
134 provided their written informed consent before completing the survey. The survey was conducted
135 on the Qualtrics platform. Before beginning the experiment, participants were first required to
136 answer three basic logic questions to deter any non-human agents from responding to the survey.
137 Participants were then asked to answer each question as honestly as they could and without
138 cheating (e.g., by looking up any of the questions online). Forty-one individuals who reported
139 cheating at the task were excluded from the analyses; analyses were conducted on the data of the
140 remaining 459 participants.

141 Participants completed 100 trials each, with each trial comprising one statement that was
142 either true or false. Half the statements at each level of difficulty were true, and the other half
143 were false. Participants were asked to provide their predictions about (a) whether the statement
144 was more likely to be true or false, (b) what percentage of other forecasters would predict the
145 statement to be true, (c) the probability that the statement was true, and (d) what the average
146 probability estimated by other forecasters would be. Participants who provided votes that were
147 inconsistent with their probability forecasts (i.e., voting “true” but predicting a probability <50%
148 of the statement being true, or voting “false” but predicting a probability >50% of the statement
149 being true) were excluded from the analysis from that particular question. Each participant saw 20
150 statements from each level of difficulty, and statements were presented in one of five randomized
151 orders. Participants who took part in any of our previous experiments were excluded from
152 participating.

153 **3.2 Alternative algorithms and planned analyses**

154 Our main algorithm of interest is the meta-probability weighting (MPW) algorithm, which
155 weights forecasters’ probability forecasts by the normalized absolute difference between their
156 probability forecasts and their meta-predictions about the average probability forecasted by

157 others. Our comparison set also includes three other approaches from the literature: the simple
158 average, the p''_{cs} aggregator [8], and the minimal pivoting procedure [13]. The details of each
159 aggregation approach used are shown in Table 1.

160 The p''_{cs} aggregator of Satopää, Pemantle, and Ungar [8] was designed to correct for the
161 conservative bias that is consistently seen in probabilistic forecasting [9, 10, 11, 12]. As discussed
162 in detail in [8], the algorithm is informed by a *partial information* framework that models the
163 amount of information overlap in forecasters. While estimation of the parameters of the full
164 model is possible with records of forecasters' past performance, a simpler model—the p''_{cs}
165 aggregator—can be applied by assuming that the information available to forecasters is
166 compound symmetric, such that forecasters' information sets have the same size and the amount
167 of pairwise overlap is constant. Assuming compound symmetry, the p''_{cs} aggregator is able to
168 estimate the amount of overlap in information between forecasters and therefore correct for this
169 overlap by extremizing probability forecasts such that forecasts of low probabilities are shifted
170 closer to 0 and forecasts of high probabilities are shifted closer to 1. Empirically, the authors
171 found that the p''_{cs} aggregator outperformed simple averaging and also both log-odds and probit
172 aggregators on a large dataset of real-world geopolitical forecasting problems from the ACE
173 forecasting tournament.

174 Palley and Soll [13] utilized a different approach, the *minimal pivoting* procedure, to
175 correct for bias in the aggregated crowd forecast due to the sharing of information by adjusting
176 the average forecast using forecasters' meta-predictions about the average forecast of others. The
177 authors showed that the optimal correction (or *pivot*) for this bias depends on the structure of
178 shared information between forecasters. For example, the optimal amount of pivoting for a crowd
179 of laypeople will differ to the optimal amount of pivoting for a crowd of experts. As the structure
180 of shared information for a given problem may be unknown to the decision-maker beforehand, the
181 authors proposed the use of a minimal pivoting procedure, which provides a conservative

Table 1. Details of each aggregation approach used. The name, formula, and description for each probabilistic aggregation approach used in this paper. The notation for each aggregation approach is explained in the main text above, excluding the p''_{cs} aggregator, for which, due to its complexity, we refer readers to the original paper by Satopää et al. [8].

Aggregation approach	Formula	Description
Simple average	$T_{\mu}(X_k) = \sum_{i=1}^{N_k} \frac{P_{i,k}}{N_k}$	Simple unweighted average of all individual forecasts in the crowd.
p''_{cs}	$T_{p''_{cs}}(X_k) = \Phi \left(\frac{\frac{1}{(N-1)\lambda + 1} \sum_{i=1}^N X_{B_i}}{\sqrt{1 - \frac{N\delta}{(N-1)\lambda + 1}}} \right)$	Revealed Aggregator for the Gaussian Model under compound symmetry – see Satopää et al. [8] for details.
Minimal Pivoting	$T_{MP}(X_k) = \sum_{i=1}^{N_k} \frac{P_{i,k} + (P_{i,k} - M_{i,k}^P)}{N_k}$	Simple average corrected by the minimal pivoting procedure [13].
Meta-probability Weighting (MPW)	$T_{MPW}(X_k) = \sum_{i=1}^{N_k} \frac{ P_{i,k} - M_{i,k}^P P_{i,k}}{\sum_{j=1}^{N_k} P_{j,k} - M_{j,k}^P }$	Weighted average of forecasters' probability forecasts, where weights are calculated from the normalized absolute difference between their probability forecasts and their meta-predictions about the average probability forecasted by others.

187 correction relative to the optimal pivoting procedure when the information structure is known.
 188 The authors tested the minimal pivoting procedure across four studies and found that minimal
 189 pivoting outperformed simple averaging on both a cost-estimation task and sports prediction
 190 problems.

191 While we could have applied these aggregation approaches directly, many previous
 192 studies have highlighted the consistent need for extremisation in the probabilistic forecasting
 193 domain [9, 10, 11, 14, 15, 16]. We therefore considered two versions of each algorithm: the
 194 standard version and a version augmented using the extremisation function used by Baron et al.
 195 [9] and others before them [10, 11]:

$$t(p) = \frac{p^a}{p^a + (1 - p)^a} \quad (2)$$

196 where p is the original aggregated probability forecast, $t(p)$ is the recalibrated probability, and a is
 197 the recalibration parameter, which determines the strength of the transformation. This function
 198 extremises probability forecasts when $a > 1$ and anti-extremises when $0 < a < 1$. Baron et al. [9]
 199 conducted a large-scale study where over 2,000 people were asked to estimate the probabilities of
 200 outcomes to international events such as political elections occurring a future date. Baron et al. [9]
 201 found that the optimal parameter value for this function was approximately $a = 2.5$ in crowds
 202 containing expert forecasters, who, on average, were under-confident and therefore needed to be
 203 extremised to become optimally calibrated. For this reason, we selected this parameter value in
 204 advance and applied it to each aggregation approach. Extremisation improved forecasts for the
 205 simple average, MPW algorithm, and minimal pivoting procedure, but not for the p''_{cs} aggregator,
 206 which already produced extremised forecasts [8]. In our results, we report the comparison
 207 between the extremised version of the MPW algorithm and both the standard and extremised
 208 versions of each other aggregation approach.

In line with Budescu and Chen [6] and Chen et al. [17], we compare the performance of the MPW algorithm and other probabilistic aggregation approaches using a transformed Brier score:

$$S_i = 100 - 100 \sum_{k=1}^K \frac{(D(o_k) - T(X_{i,k}))^2}{K}, \quad (3)$$

where S_i is the score of the i^{th} forecaster (or algorithm) averaged across K total events, $D(o_k)$ is the outcome variable for the k^{th} event (equals 1 if the event is true and 0 if false), and $T(X_{i,k})$ is the probability assigned to that outcome being true by that forecaster (or algorithm). This scoring rule has a straightforward interpretation where scores range from 0 to 100, with 100 being a perfect forecast over all events. Importantly, this linear transformation of the Brier score retains the same functional form as the original and is strictly proper [18]. Strictly proper scoring rules are conventional measures of performance in probabilistic forecasting and are useful because they ensure that performance of the probability forecasts, measured as some sort of score, is optimized only by forecasts of the true probability. The use of scoring rules in assessing forecasts thus encourages forecasters to be careful and truthful in making their forecasts, in order to maximize their score.

We assess statistical significance between predictions of different aggregation approaches using 95% confidence intervals (CIs), which indicate, firstly, a significance difference when the null hypothesis value ($H_0 = 0$) is not contained within the interval, and secondly, a plausible range for the size of the effect. We compute 95% confidence intervals for paired differences in transformed Brier score between the MPW algorithm and each other approach using the empirical bias-corrected and accelerated (BCa) bootstrap [19] using 10,000 bootstrap samples. Confidence intervals were computed using the standard *bootci* function in the MATLAB program. We have included the experimental data and MATLAB code for the analyses and plots from this paper in the supplementary information files.

232 4. Results

233 Fig 1 shows the mean performance for each aggregation approach across the 500
234 problems. After extremisation, the MPW algorithm generated significantly better predictions
235 overall than: the standard mean individual by 14.22 points (bootstrap 95% CIs for paired mean
236 difference in score: [13.04, 15.36]), the extremised mean individual by 18.20 points (95% CI:
237 [16.86, 19.57]), the standard simple average by 6.64 points (95% CI: [5.61, 7.63]), the extremised
238 simple average by 6.23 points (95% CI: [4.91, 7.62]), the standard p''_{cs} aggregator by 5.04 points
239 (95% CI: [3.83, 6.44]), the extremised p''_{cs} aggregator by 7.24 points (95% CI: [5.40, 9.33]), the
240 standard minimal pivoting procedure by 4.21 points (95% CI: [3.37, 4.98]), and the extremised
241 minimal pivoting procedure by 3.43 points (95% CI: [2.47, 4.47]). The MPW algorithm was
242 therefore highly effective at generating probabilistic forecasts across a range of low-difficulty to
243 high-difficulty decision problems.

244 [Figure 1 Here]

245 **Fig 1. Overall performance of the standard vs. extremised versions of each aggregation**
246 **approach.** The mean transformed Brier score over a total of 500 US grade school problems
247 spanning five levels of difficulty. Error bars indicate the standard error. The standard version of
248 each approach generates probabilistic forecasts according to their formulae shown in Table 1. The
249 extremised version of each approach transforms these predictions using a simple extremisation
250 function [9]. The extremised MPW algorithm significantly outperforms both the standard and
251 extremised versions of every other aggregation approach.

253 We examined whether the improvement offered by the MPW algorithm over simple
254 averaging varied across different problem difficulties. As the MPW algorithm leverages latent
255 expertise, we would expect it to offer the greatest improvement over simple averaging on
256 moderate-difficulty and high-difficulty forecasting problems, where the crowd is likely comprised
257 of both experts and novices. Fig 2 shows the mean performance of the best-performing versions
258 of each aggregation approach separately for each of the five difficulty levels. Table 2 shows the
259 mean difference in transformed Brier score between the extremised MPW algorithm and each

other approach for each difficulty. While the extremised MPW algorithm outperformed all other approaches on the problem sets from difficulties 2 to 5, this improvement was only significant for all comparisons from difficulty 2, 3, and 5.

[Figure 2 Here]

Fig 2. Performance of each aggregation approach on each level of difficulty. The mean transformed Brier score for each level of difficulty of US grade school problems. Error bars indicate the standard error. The extremised MPW algorithm (blue bar) outperforms the best-performing version of all other aggregation approaches on problems from difficulties 2 to 5. The 95% CIs for mean difference in score between the extremised MPW algorithm and each other aggregation approach is shown in Table 2.

270
271
272
273

Table 2. 95% Confidence intervals for the mean difference in the transformed Brier score between the extremised MPW algorithm and the standard and extremised versions of each other aggregation approach. Asterisks indicate where the difference in score was statistically significant at the $\alpha = .05$ level according to the paired mean difference in transformed Brier score using the BCa bootstrap [19].

Aggregation approach	Version	Difficulty 1	Difficulty 2	Difficulty 3	Difficulty 4	Difficulty 5
Mean individual	Standard	[11.22, 15.58]*	[14.44, 19.65]*	[12.79, 18.19]*	[7.93, 13.89]*	[12.04, 16.79]*
	Extremised	[12.97, 18.48]*	[17.32, 23.78]*	[16.90, 23.08]*	[13.16, 19.23]*	[16.49, 22.04]*
Simple average	Standard	[3.62, 6.95]*	[6.25, 10.44]*	[5.63, 10.23]*	[1.06, 6.78]*	[5.24, 9.59]*
	Extremised	[-0.85, 3.03]	[3.54, 9.75]*	[5.86, 11.57]*	[3.23, 9.38]*	[5.79, 12.72]*
p''_{cs}	Standard	[0.47, 6.11]*	[3.02, 8.98]*	[3.66, 8.49]*	[0.53, 6.75]*	[5.42, 12.03]*
	Extremised	[-1.01, 6.63]	[2.32, 11.16]*	[5.07, 13.25]*	[3.73, 12.41]*	[7.69, 18.29]*
Minimal pivoting	Standard	[1.44, 3.94]*	[3.61, 6.80]*	[3.20, 6.89]*	[-0.17, 4.53]	[3.68, 7.22]*
	Extremised	[-2.23, 0.47]	[0.80, 5.11]*	[2.88, 6.88]*	[1.87, 6.50]*	[3.75, 9.47]*

* indicates where $p < .05$

274

The extremised MPW algorithm performed particularly well relative to other approaches on the problems in the highest difficulty level. For example, the extremised MPW algorithm outperformed simple averaging by approximately 9 points in score, which was approximately three times as large an improvement compared to that offered by the next best approach, the extremised minimal pivoting procedure. Consistent with our predictions, the extremised MPW algorithm also performed equally well as other aggregation approaches on the lowest difficulty level. Our empirical findings are thus highly consistent with the predictions of our theoretical model. These results provide strong evidence for the MPW algorithm’s mechanism to leverage latent crowd expertise, a mechanism that is most effective on moderate-difficulty to high-difficulty forecasting problems where the crowd is likely to be comprised of both experts and novices.

One explanation for our results is that the parameter values chosen for the extremisation function were simply better suited for the extremised MPW algorithm than these other aggregation approaches. Although we based our choice of parameter values from previous results from other authors [9], it could be the case that these values were simply optimized for the MPW algorithm and not the other aggregation approaches. To address this concern, we conducted additional post-hoc analyses to investigate whether optimally recalibrating these other aggregation approaches could allow them to outperform the extremised MPW algorithm. We optimally recalibrated each other aggregation approach using that approach’s responses to other forecasting problems (i.e., using cross-validation when past performance is known). For each approach, we used leave-one-out cross-validation to estimate the optimal parameter (a) in the recalibration function adapted from Baron et al. [9]. For each training set, we tested a range of values for a from 0 to 10 in increments of 0.01 and selected the value that maximized the score of that approach, which we then applied to the training event. We repeated this process separately for each of the 500 questions in the dataset, and for each of the five aggregation approaches. For

statistical inference, we used the BCa bootstrap [19] with 10,000 bootstrap samples to compute 95% CIs for the mean paired difference in score between aggregation approaches.

Fig 3 shows the performance of these other aggregation approaches once they have been optimally recalibrated. While optimizing the recalibration function for these other approaches improved their performance, the extremised MPW algorithm, which was not optimally recalibrated, still offered significantly better predictions than any other approach. Comparing the mean performance of the fixed version of the extremised MPW algorithm to the other optimally recalibrated approaches, we find that the extremised MPW algorithm outperforms each other approach even when they have been optimally recalibrated. The fixed extremised MPW algorithm scored higher than the optimally-recalibrated simple average by 5.79 points (95% CI: [4.66, 6.94]), the optimally-recalibrated p''_{cs} aggregator by 5.19 points (95% CI: [3.91, 6.65]), and the optimally-recalibrated minimal pivoting procedure by 3.15 points (95% CI: [2.33, 4.01]).

[Figure 3 Here]

Fig 3. Performance of each aggregation approach using cross-validated recalibration parameters. This figure shows the mean performance of each approach using the fixed parameter value $a = 2.5$ (orange bars) vs. optimal recalibration parameters estimated via cross-validation (blue bars). Error bars show the standard error.

In the S2 Appendix, we also conducted a post-hoc analysis where we compared the extremized version of each aggregation approach included in this paper to a dataset containing forecasts about NCAA men’s basketball games that was collected by Palley & Soll [13]. In this dataset we find no significant difference between the performance of the extremised MPW algorithm, the extremised minimal pivoting mechanism, the p''_{cs} aggregator, or the extremised simple average. The dataset does not appear to have any experts in it, which may account for the similar prediction of all four methods.

5. Discussion

In the current paper, we have developed a novel algorithm for leveraging forecasters' expertise using forecasters' meta-predictions about what other forecasters would predict. The extremised MPW algorithm allows decision-makers to generate accurate probabilistic predictions even when the forecasters' past performance is unavailable. The extremised MPW algorithm is also computationally simple, which may be appealing to decision-makers that are unfamiliar with more-sophisticated aggregation approaches that require structural estimation of latent parameters [20]. While previous research have demonstrated how meta-predictions can be used to correct for crowd biases [7], or used to identify the structure and extent of shared information in the crowd [13], no studies to date have shown that forecasters' meta-predictions can be used to derive weights that quantify latent expertise. The extremised MPW algorithm is therefore theoretically distinct from existing approaches such as the p''_{cs} aggregator [8] and the minimal pivoting procedure [13], which seek to improve forecasts by modelling and correcting for the overlap in information between forecasters.

The current paper provides a valuable contribution in demonstrating that this empirical quantity can be used to produce probabilistic forecasts that outperform existing aggregation approaches in the literature. In particular, the extremised MPW algorithm outperforms other existing aggregation approaches that can be applied on forecasting problems where the forecasters' past performance is unknown: simple averaging, the p''_{cs} aggregator [8], and the minimal pivoting procedure [13]. Relative to these other approaches, the extremised MPW algorithm performs particularly well for the more difficult forecasting problems, where leveraging latent expertise is likely to be most effective. Decision-makers who are faced with difficult forecasting problems may therefore find the extremised MPW algorithm an attractive alternative over existing aggregation approaches.

350 **Data Availability Statement**

351 All data analysed during this study are included in this published article (and its
352 Supplementary Information files), excluding the data analysed in the S2 Appendix, which is
353 available from the original authors (see Palley & Soll [13]).
354

355 **References**

356 [1] Galton F. Vox populi (The wisdom of crowds). *Nature*. 1907;75(7):450–451.

357 [2] Armstrong JS. Principles of forecasting: a handbook for researchers and practitioners.
358 vol. 30. Springer Science & Business Media; 2001.

359 [3] Cooke RM. Experts in uncertainty: opinion and subjective probability in science. Oxford
360 University Press on Demand; 1991.

361 [4] Cooke RM, Goossens LH. Procedures guide for structural expert judgement in accident
362 consequence modelling. *Radiation Protection Dosimetry*. 2000;90(3):303–309.

363 [5] Mellers B, Baker JD, Chen E, Mandel DR, Tetlock PE. How generalizable is good
364 judgment? A multi-task, multi-benchmark study. *Judgment and Decision making*.
365 2017;12(4):369–382.

366 [6] Budescu DV, Chen E. Identifying expertise to extract the wisdom of crowds. *Management
367 Science*. 2015;61(2):267–280.

368 [7] Prelec D, Seung HS, McCoy J. A solution to the single-question crowd wisdom problem.
369 *Nature*. 2017;541(7638):532.

370 [8] Satopää VA, Pemantle R, Ungar LH. Modeling probability forecasts via information
371 diversity. *Journal of the American Statistical Association*. 2016;111(516):1623–1633.

372 [9] Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH. Two reasons to make aggregated
373 probability forecasts more extreme. *Decision Analysis*. 2014;11(2):133–145.

374 [10] Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS. Forecast aggregation via
375 recalibration. *Machine learning*. 2014;95(3):261–289.

376 [11] Shlomi Y, Wallsten TS. Subjective recalibration of advisors’ probability estimates.
377 *Psychonomic bulletin & review*. 2010;17(4):492–498.

378 [12] Dana J, Atanasov P, Tetlock P, Mellers B. Are markets more accurate than polls? The
379 surprising informational value of “just asking”. *Judgment and Decision Making*.
380 2019;14(2):135–147.

381 [13] Palley AB, Soll JB. Extracting the Wisdom of Crowds When Information Is Shared.
382 *Management Science*. 2019;65(5):2291–2309.

383 [14] Ranjan R, Gneiting T. Combining probability forecasts. *Journal of the Royal Statistical*
384 *Society: Series B (Statistical Methodology)*. 2010;72(1):71–91.

385 [15] Erev I, Wallsten TS, Budescu DV. Simultaneous over-and underconfidence: The role of
386 error in judgment processes. *Psychological review*. 1994;101(3):519.

387 [16] Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH. Combining multiple
388 probability predictions using a simple logit model. *International Journal of Forecasting*.
389 2014;30(2):344–356.

390 [17] Chen E, Budescu DV, Lakshmikanth SK, Mellers BA, Tetlock PE. Validating the
391 contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*.
392 2016;13(2):128–152.

393 [18] Murphy AH, Winkler RL. Scoring rules in probability assessment and evaluation. *Acta*
394 *psychologica*. 1970;34:273–286.

395 [19] Efron B. Better bootstrap confidence intervals. *Journal of the American statistical*
396 *Association*. 1987;82(397):171–185.

397 [20] McCoy J, Prelec D. A statistical model for aggregating judgments by incorporating peer
398 predictions. *arXiv preprint arXiv:170304778*. 2017;.

399 [21] Blackwell D. Equivalent comparisons of experiments. The Annals of Mathematical
400 Statistics. 1953;p. 265–272.

401 [22] Blackwell D, Girshick MA. Theory of games and statistical decisions. Courier
402 Corporation; 1979.

403 [23] Marschak J, Miyasawa K. Economic comparability of information systems. International
404 Economic Review. 1968;9(2):137–174.

405 [24] Marschak J, Radner R. Economic Theory of Teams (Cowles Foundation Monograph 22).
406 Yale University Press, New Haven, CT; 1972.

407 [25] Blackwell D. Comparison of Experiments. In: Proceedings of the Second Berkeley
408 Symposium on Mathematical Statistics and Probability. The Regents of the University of
409 California; 1951. .

410

411 **Acknowledgments**

412 We wish to thank Asa Palley and Jack Soll for kindly sharing with us the experimental
413 data that they had collected [13].

414

415 **Author Contributions**

416 All authors were involved in the study design, data collection, data analysis, interpretation,
417 and writing the manuscript.

418

419 **Competing Interests Statement**

420 The authors declare no competing interests.

421 **Supplementary Information**

- 422 • **S1 Appendix. Theory appendix for understanding how the MPW algorithm**
423 **leverages expertise.**
- 424 • **S2 Appendix. Testing the MPW algorithm on Palley & Soll (2018)’s NCAA Men’s**
425 **Basketball dataset.**
- 426 • **S1 File: Experimental questions, participant responses, and analysis code.**
- 427 • **References [21, 22, 23, 24, 25]**

Appendix S1. Theory appendix for understanding how the MPW algorithm leverages expertise

In this appendix, we investigate the conditions under which the meta-probability weighting (MPW) algorithm is able to leverage expertise by weighting the probabilistic predictions of experts more than novices. We show that under very general conditions, an individual with a more informative signal about the true state will be weighted more heavily by the algorithm than an individual with a less informative signal. This implies that from an ex-ante perspective, an individual who has access to a more informative information system will have a higher expected weight than an individual with a less informative information system.

1. Preliminaries

As is common in the information economics literature, we model expertise by considering an environment in which individuals have access to an information system (often called an experiment) in which they receive a signal that they use to update an initial prior belief [21–24]. Experts and novices are distinguished by the informativeness of their information system but are identical in all other dimensions.

We consider a Bayesian model in which a crowd of forecasters share a common prior $p(T)$ that an event is true. Each forecaster receives a private signal S , that is a random variable taking on real value realizations in the set $\{s_1, \dots, s_m\} \cup \{s_\emptyset\}$ where $0 \leq s_1 < s_2 < \dots < s_m \leq 1$ and $s_1 < s_\emptyset < s_m$. As our outcome space is binary, it is without loss of generality that we normalize the signals so that their value is equal to the posterior belief that an event is true. That is, $s_k := p(T|s_k)$. We let $s_\emptyset$ represent the case where an individual receives an uninformative signal so that $s_\emptyset := p(T)$.

The potential signals that an individual can receive is based on the information service that each forecaster has access to. We assume that there are two alternative information services — one for experts and one for the novices — with likelihood matrices $[Q_{oj}^E]_{2 \times (m+1)}$ and $[Q_{oj}^N]_{2 \times (m+1)}$. Each element of the first row of Q^E and Q^N represents the probability that the signal is s_j given the outcome is $o = T$. Likewise, each element of the second row of Q^E and Q^N represent the probability that the signal is s_j given the outcome is $o = F$. For ease, we will denote the first row elements with T and the second row elements with F . Thus $Q_{Tj}^E := Q_{1j}^E = p(s_j|T)$ while $Q_{Fj}^E := Q_{2j}^E = p(s_j|F)$.

We note two important features of an information service. First, an information service acts as a transition matrix from a state of nature to a signal and thus $\sum_j Q_{oj}^\tau = 1$ for each row $o \in \{T, F\}$ and information service $\tau \in \{N, E\}$. Second, upon receiving a message from information service τ , agents revise their priors using Bayes rule. For any signal that occurs with positive probability (i.e., where $Q_{Tj}^\tau + Q_{Fj}^\tau > 0$), the posterior belief that the event is true is given by

$$p(T|s_j) = \frac{p(T)Q_{Tj}^\tau}{p(T)Q_{Tj}^\tau + p(F)Q_{Fj}^\tau}.$$

By construction, this is equal to s_j for all signals that occur with positive probability.

We assume that the proportion of experts in the crowd is known to all parties and given by $\theta \in [0, 1]$. We also assume that the properties of Q^E and Q^N are common knowledge. We make two additional assumptions regarding the information services used by novices and experts:

Assumption 1 *Information service Q^E is more informative than information service Q^N : there exists a non-negative stochastic matrix $Z = [Z_{ki}]_{(m+1) \times (m+1)}$ such that*

$$Q^N = Q^E Z.$$

Assumption 1 says that when Q^E is more informative than Q^N , $Q_{oi}^N = \sum_k Q_{ok}^E Z_{ki}$. As we are multiplying across the rows of Q^E , we can interpret Z_{ki} as the conditional probability that when message k is received by Q^E , message i was received by Q^N . Thus $Z_{ki} = p(s_i|s_k)$ and Q^E is more informative than Q^N if it is possible to garble the signals of Q^E and generate Q^N . Note that Z is a non-negative stochastic matrix with $\sum_i Z_{ki} = 1$.

Assumption 2 *Experts and Novices draw independent signals: for a signal s_i from Q^N and a signal s_k from Q^E ,*

$$p(s_i, s_k) = p(s_i|T)p(s_k|T)p(T) + p(s_i|F)p(s_k|F)p(F)$$

Assumption 2 implies that the information services are ranked but that signals from the two information services are independent once we condition for the state. Thus, for any signal s_i drawn from $Q^t, t \in \{N, E\}$, and any signal s_j drawn from $Q^\tau, \tau \in \{N, E\}$,

$$p(s_i|s_j) = \frac{p(s_i, s_j)}{p(s_j)} = \frac{p(s_i|T)p(s_j|T)p(T) + p(s_i|F)p(s_j|F)p(F)}{p(s_j)}.$$

Rearranging Bayes Rule, it is the case that:

$$\frac{p(T)p(s_j|T)}{p(s_j)} = p(T|s_j) = s_j$$

and thus

$$p(s_i|s_j) = p(s_i|T)s_j + p(s_i|F)(1 - s_j) = Q_{Ti}^t s_j + Q_{Fi}^t (1 - s_j). \quad (1)$$

We note that Assumption 2 is also sufficient for the monotone likelihood ratio property (MLRP) to hold for signals between any two information services. This property implies that when an individual receives a high signal, he believes that other forecasters are also more likely to receive a high signal.

2. The Expected Contribution of Experts and Novices

We now turn to the question of how the MPW algorithm weights experts and novices. A forecaster with signal s_k will make a probabilistic forecast of s_k . Thus, given an outcome state o ,

the expected prediction from information service Q^t is given by

$$P(Q^t|o) = \sum_{\{i|s_i \geq 0\}} Q_{oi}^t s_i.$$

Aggregating over both information systems, the expected prediction of the population in state o is given by

$$P(\theta|o) = \sum_{\{i|s_i \geq 0\}} \theta P(Q^E|o) + (1 - \theta) P(Q^N|o).$$

In the absence of any information service, the probabilistic forecast of each individual would be $s_\emptyset$. By the law of total expectations, the posteriors are a mean-preserving spread of the prior, and thus we have

$$s_\emptyset = s_\emptyset P(Q^\tau|T) + (1 - s_\emptyset) P(Q^\tau|F).$$

for $\tau \in \{E, N\}$. This also implies that

$$s_\emptyset = s_\emptyset P(\theta|T) + (1 - s_\emptyset) P(\theta|F)$$

and that

$$P(\theta|F) = \frac{s_\emptyset}{1 - s_\emptyset} [1 - P(\theta|T)]. \quad (2)$$

A forecaster with signal s_k 's meta-prediction about the others is equal to

$$M(\theta|s_k) = s_k P(\theta|T) + (1 - s_k) P(\theta|F).$$

Substituting in for $P(\theta|F)$ using (2), the meta-prediction of an individual with signal s_k can be expressed as

$$M(\theta|s_k) = s_k P(\theta|T) + (1 - s_k) \frac{s_\emptyset}{1 - s_\emptyset} [1 - P(\theta|T)].$$

In the MPW algorithm, the weight of an individual is based on the difference between the individual's prediction and meta-prediction. For an individual with signal s_k ,

$$s_k - M(\theta|s_k) = s_k - s_k P(\theta|T) - (1 - s_k) \frac{s_\emptyset}{1 - s_\emptyset} [1 - P(\theta|T)]$$

or, equivalently:

$$s_k - M(\theta|s_k) = \frac{s_k - s_\emptyset}{1 - s_\emptyset} [1 - P(\theta|T)]. \quad (3)$$

Note first that the difference between an individual's signal and his or her meta-prediction is zero at $s_\emptyset$ and is linearly increasing in s_k . This feature implies that the weight of an individual with signal s_k , proportional to $|s_k - M(\theta|s_k)|$, is directly related to the informativeness of the posterior that an individual holds relative to the prior. Thus, individuals with more informative posteriors (an ex-post notion of expertise) will be weighted proportionally more than individuals with less informative posteriors.

Aggregating over all possible signals that an individual might receive, the expected weight of an individual in information service Q^t is proportional to:

$$\mathbb{E}[w|Q^t] = \sum_{\{k|s_k>0\}} |s_k - M(\theta|s_k)|[s_{\emptyset}Q_{Tk}^t + (1 - s_{\emptyset})Q_{Fk}^t].$$

By Assumption 1, the information system of the expert is a mean-preserving spread of the signals of the novices. Noting that individuals with posteriors farther away from their prior will have a larger weight, it is possible to use Blackwell's theorem to show the following:

Proposition 1 *The expected weight of an expert is greater than the expected weight of a novice.*

We note that Proposition 1 does not rely on us having only two types of agents in the population and can be readily extended to an arbitrary number of information systems that are ranked by informativeness. Thus, the results here are general and are likely to hold in a wide variety of problems. We also note that, the MPW algorithm weights experts more than novices if we use an ex-ante notion of expertise based on informativeness of information systems or an ex-post notion of expertise based on the difference between an individual's posterior and their prior.

3. Proof of Proposition 1

Proof of Proposition 1: We begin this proof of Proposition 1 by stating Blackwell's Theorem [25]:

Blackwell's Theorem *For information service Q^E to be more informative than Q^N it is necessary and sufficient that the value of information in service Q^E is greater than the value of information in service Q^N for all sets of terminal actions, all utility functions, and all prior beliefs.*

By Assumption 1, Q^E is more informative than Q^N . Let the action set $a \in \{T, F\}$ correspond to voting on whether an answer is true or false, and consider a utility function $U(a, o)$ that maps actions and states of the world into outcomes. Let $U(T, T) = \frac{1}{1-s_\emptyset}[1-P(\theta|T)] - \frac{s_\emptyset}{1-s_\emptyset}[1-P(\theta|T)]$, $U(F, F) = \frac{1}{1-s_\emptyset}[1-P(\theta|t)] - [1-P(\theta|T)]$, $U(T, F) = -\frac{s_\emptyset}{1-s_\emptyset}[1-P(\theta|T)]$, and $U(F, T) = -[1-P(\theta|T)]$. Given a signal s_i , expected utility is maximized by choosing $a = T$ when $s_i \geq s_\emptyset$ and $a = F$ otherwise. The expected utility of this strategy when the posterior is less than $s_\emptyset$ is given by:

$$\begin{aligned} \mathbb{E}[U(Q^t|s_i < s_\emptyset)] &= (1-s_i)U(F, F) + s_iU(F, T) \\ &= \frac{1-s_i}{1-s_\emptyset}[1-P(\theta|T)] - (1-s_i)[1-P(\theta|T)] - s_i[1-P(\theta|T)] \\ &= \frac{1-s_i}{1-s_\emptyset}[1-P(\theta|T)] - [1-P(\theta|T)] \\ &= \frac{s_\emptyset - s_i}{1-s_\emptyset}[1-P(\theta|T)]. \end{aligned}$$

Likewise, the expected utility of this strategy given a posterior greater than $s_\emptyset$ is

$$\begin{aligned} \mathbb{E}[U(Q^t|s_i \geq s_\emptyset)] &= s_iU(T, T) + (1-s_i)U(T, F) \\ &= \frac{s_i}{1-s_\emptyset}[1-P(\theta|T)] - s_i\frac{s_\emptyset}{1-s_\emptyset}[1-P(\theta|T)] - (1-s_i)\frac{s_\emptyset}{1-s_\emptyset}[1-P(\theta|T)] \\ &= \frac{s_i}{1-s_\emptyset}[1-P(\theta|T)] - \frac{s_\emptyset}{1-s_\emptyset}[1-P(\theta|T)] \\ &= \frac{s_i - s_\emptyset}{1-s_\emptyset}[1-P(\theta|T)]. \end{aligned}$$

Thus, we can express the expected utility of an individual with signal s_i as

$$\mathbb{E}[U(Q^t|s_i)] = \frac{|s_i - s_\emptyset|}{1-s_\emptyset}[1-P(\theta|T)]. \quad (4)$$

By Blackwell's theorem, the expected utility of information service Q^E is higher than the expected value of information service Q^N for any utility function and any prior belief. Using an initial prior of $P(T) = s_\emptyset$, this implies,

$$\mathbb{E}[U(Q^E)] \geq \mathbb{E}[U(Q^N)]$$

By the law of iterated expectations:

$$\begin{aligned}\mathbb{E}[U(Q^t)] &= \sum_{\{i|s_i \geq 0\}} \mathbb{E}[U(Q^t|s_i)]p(s_i) \\ &= \sum_{\{i|s_i \geq 0\}} \mathbb{E}[U(Q^t|s_i)][s_{\emptyset}Q_{Ti}^t + (1 - s_{\emptyset})Q_{Fi}^t].\end{aligned}$$

Using the result in (4) above, this implies that

$$\mathbb{E}[U(Q^t)] = \sum_{\{i|s_i \geq 0\}} \frac{|s_i - s_{\emptyset}|}{1 - s_{\emptyset}} [1 - P(\theta|T)][s_{\emptyset}Q_{Ti}^t + (1 - s_{\emptyset})Q_{Fi}^t]$$

Noting that this equation is identical to the expected weight of an individual from information sytem Q^E

$$E[w|Q^t] = \sum_{\{i|s_i \geq 0\}} \frac{|s_i - s_{\emptyset}|}{1 - s_{\emptyset}} [1 - P(\theta|T)][s_{\emptyset}Q_{Ti}^t + (1 - s_{\emptyset})Q_{Fi}^t],$$

the result that $\mathbb{E}[U(Q^E)] \geq \mathbb{E}[U(Q^N)]$ immediately implies that $E[w|Q^E] \geq E[w|Q^N]$. ■

S2 Appendix. Testing the MPW algorithm on Palley & Soll (2018)'s NCAA Men's Basketball dataset.

In the main text, we have concentrated our analyses only on data that we have collected expressly for the purposes of testing the MPW algorithm. Out of interest, we also sought to test the MPW algorithm on datasets in the literature that have been collected by other authors. In this Appendix, we examine the performance of the extremised MPW algorithm on the data from Study 4 of Palley & Soll [1]. In their study, the authors elicited participants' forecasts for the probability that each team would win across 120 different games in the early rounds of the 2014, 2015, and 2016 NCAA Division I Men's Basketball Tournament. Participants were recruited via Amazon Mechanical Turk and completed the experiment via an online web survey.

The responses in their dataset in which we are interested are the (1) the team that the participant predicts would win a particular match, (2) the participants forecast for the probability that their selected team would win, and (3) their estimate of the average probability forecasted by other participants that the participant's selected team would win. These responses correspond to the responses that we had collected in our experiment in the main text.

We computed the predictions for each of the aggregation approaches in the main paper using the responses in this dataset. We calculated the transformed Brier score for each of the standard and extremised version of each aggregation approach from the main and plotted the mean results below. For statistical inference, we computed 95% CIs for the mean difference in transformed Brier score between each aggregation approach's forecasts using the BCa bootstrap 95% [2].

Fig S1 shows the performance of the standard and extremized version of each aggregation approach. The extremised MPW algorithm significantly outperformed the standard mean individual (95% CI for mean difference in transformed Brier score: [1.61, 6.49]), and the extremised mean individual (95% CI: [4.97, 8.70]). However, there was no significant difference in score between the extremised MPW algorithm and any other aggregation approach: the standard simple average (95% CI: [-2.72, 2.17]), the extremised simple average (95% CI: [-2.99, -0.07]), the standard p''_{cs} aggregator (95% CI: [-3.43, 0.86]), the extremised p''_{cs} aggregator (95% CI: [-2.33, 2.90]), the standard minimal pivoting procedure (95% CI: [-2.63, 1.60]), and the extremised minimal pivoting procedure (95% CI: [-2.23, 0.31]).

Although the extremised MPW algorithm was very effective for the datasets in the main text, it does not appear to improve forecasts in the NCAA forecasting domain. We suspect that this is due to a different distribution of experts and novices in the two datasets. In our US Grades dataset, there is a large dispersion in Brier scores across forecasters and clear experts and novices. As seen in Fig S2 below, which shows the distribution of mean Brier scores in the sample population, *ex post* performance is approximately normally distributed in the basketball dataset and there is no evidence for a distinct group of experts. The MPW algorithm therefore offered no significant advantage over simple averaging as there were few-to-no latent experts in the crowd for the algorithm to identify.

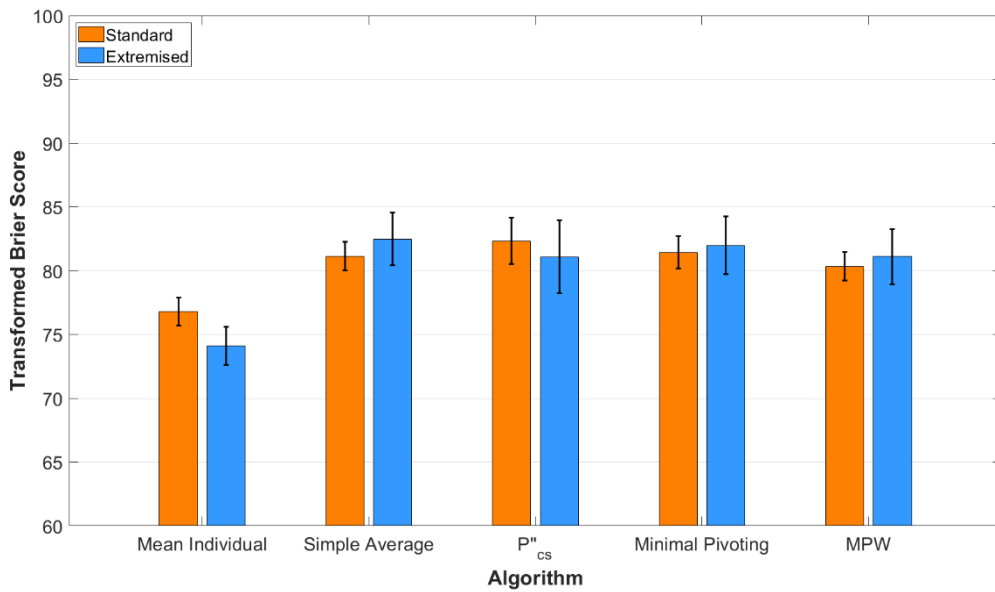


Fig S1. Overall performance of the standard vs. extremised versions of each aggregation approach in the Palley & Soll (2018) dataset. The mean transformed Brier score over a total of 120 events. Error bars indicate the standard error. The standard version of each approach generates probabilistic forecasts according to their formulae shown in Table 1. The extremised version of each approach transforms these predictions using a simple extremisation function [3]. There is no significant difference between the extremised MPW algorithm and any version of the other aggregation approaches.

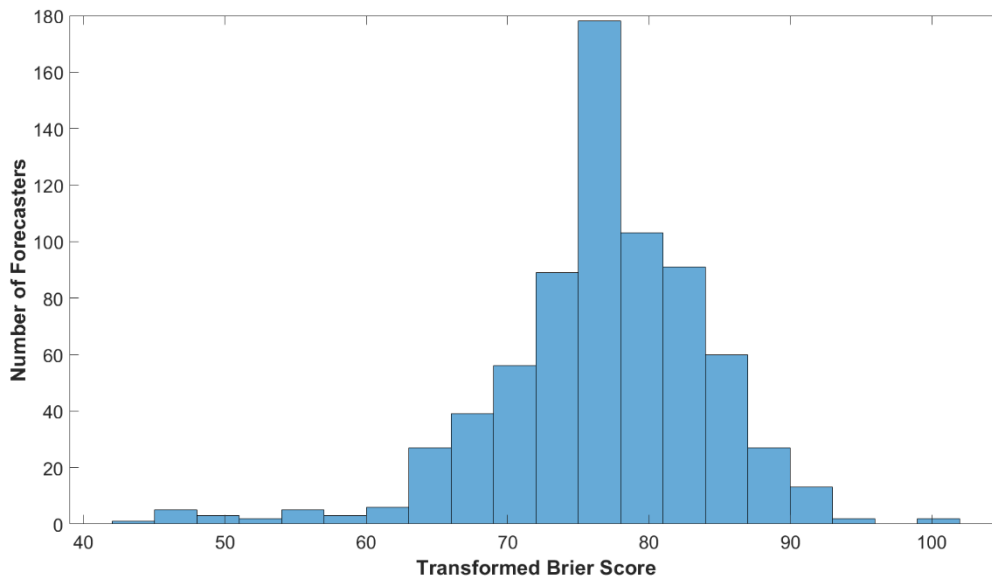


Fig S2. Distribution of forecasters' mean scores. Forecasters appear to be largely homogeneous in their *ex post* performance. There is no evidence for a distinct group of experts in the crowd.

References

- [1] Palley AB, Soll JB. Extracting the Wisdom of Crowds When Information Is Shared. *Management Science*. 2019;65(5):2291–2309.
- [2] Efron B. Better bootstrap confidence intervals. *Journal of the American statistical Association*. 1987;82(397):171–185.
- [3] Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*. 2014;11(2):133–145.

6.3 Chapter 4 Appendices

6.3.1 List of questions from each experiment in Chapter 4

Exp.	QuestionID	QuestionText	Outcome_1	Outcome_2	ActualOutcome
1	1	In the 2018 NFL draft, Mark Andrews was drafted by the	FALSE	TRUE	1
1	2	In the 2018 NFL draft, Alabama was one of the schools with no	FALSE	TRUE	1
1	3	In the 2018 NFL draft, the New York Giants were the only team	FALSE	TRUE	1
1	4	In the 2018 NFL draft, Nathan Shepard was drafted by the New	FALSE	TRUE	2
1	5	In the 2018 NFL draft, the Baltimore Ravens were the team to	FALSE	TRUE	2
1	6	In the 2018 NFL draft, the Los Angeles Rams drafted Jordan	FALSE	TRUE	1
1	7	In the 2018 NFL draft, Terrell Edmunds was drafted by the	FALSE	TRUE	2
1	8	In the 2018 NFL draft, Baker Mayfield was drafted in the	FALSE	TRUE	1
1	9	In the 2018 NFL draft, the defending Super Bowl champions	FALSE	TRUE	2
1	10	In the 2018 NFL draft, the Washington Redskins were the team	FALSE	TRUE	2
1	11	In the 2017 NFL draft, the New England Patriots had only two	FALSE	TRUE	1
1	12	In the 2017 NFL draft, more than two players were selected	FALSE	TRUE	2
1	13	In the 2017 NFL draft, the Big Ten was one of the athletic	FALSE	TRUE	1
1	14	In the 2017 NFL draft, more than two players were drafted from	FALSE	TRUE	2
1	15	In the 2017 NFL draft, more than four players were drafted from	FALSE	TRUE	1
1	16	In the 2017 NFL draft, the Chicago Bears drafted Adam Shaheen	FALSE	TRUE	2
1	17	In the 2017 NFL draft, the New York Giants were penalized in	FALSE	TRUE	2
1	18	In the 2017 NFL draft, the Minnesota Vikings had less than 12	FALSE	TRUE	1
1	19	In the 2017 NFL draft, the four special teams players drafted	FALSE	TRUE	2
1	20	In the 2017 NFL draft, the Denver Broncos drafted Trey Quinn as	FALSE	TRUE	1
1	21	In the 2016 NFL draft, the Tennessee Titans were supposed to	FALSE	TRUE	1
1	22	In the 2016 NFL draft, Carson Wentz was the only player to be	FALSE	TRUE	1
1	23	In the 2016 NFL draft, wide receiver Mortiz Bohringer was	FALSE	TRUE	1
1	24	In the 2016 NFL draft, kicker Roberto Aguayo was drafted by the	FALSE	TRUE	2
1	25	In the 2016 NFL draft, the 2015 Heisman Trophy winner, Derrick	FALSE	TRUE	1
1	26	In the 2016 NFL draft, Ohio State University had the most	FALSE	TRUE	2
1	27	In the 2016 NFL draft, more than two teams were forced to	FALSE	TRUE	2
1	28	In the 2016 NFL draft, Rico Gathers was drafted by the Oakland	FALSE	TRUE	1
1	29	In the 2016 NFL draft, David Onyemata was drafted by the New	FALSE	TRUE	2
1	30	In the 2016 NFL draft, the Tennessee Titans drafted Kalan Reed	FALSE	TRUE	2
1	31	In NFL rules, a player performing a chop block will be penalized	FALSE	TRUE	1
1	32	In NFL rules, a player who wears illegal equipment is to be	FALSE	TRUE	1
1	33	In NFL rules, a delay of game penalty at the start of either half is	FALSE	TRUE	1

1	34	In NFL rules, the penalty for attempting to use more than 3	FALSE	TRUE	2
1	35	In NFL rules, when holding is committed by the defense, the	FALSE	TRUE	2
1	36	In NFL rules, the penalty for the first onside kick going out of	FALSE	TRUE	2
1	37	In NFL rules, pass interference and neutral zone infractions are	FALSE	TRUE	2
1	38	In NFL rules, the penalty for leaping is 5 yards when it is	FALSE	TRUE	1
1	39	In NFL rules, when an intentional grounding penalty is called at	FALSE	TRUE	2
1	40	In NFL rules, the penalty for any action which delays the next	FALSE	TRUE	1
1	41	In NFL, a sack is when a quarterback or other player on offense	FALSE	TRUE	1
1	42	In NFL, the quarterback is permitted to change the play if he	FALSE	TRUE	2
1	43	In NFL, a "three and out" refers to when three wide receivers set	FALSE	TRUE	1
1	44	In NFL, there 7 rounds in the draft, with the worst-performing NFL team from the previous year choosing first, and the best-	FALSE	TRUE	2
1	45	In NFL, a "Hail Mary" is a play in which the receivers are all sent	FALSE	TRUE	2
1	46	In NFL, a "two-point conversion" is a play a team attempts instead of kicking a one-point conversion immediately after it	FALSE	TRUE	2
1	47	In NFL, Vince Lombardi led the Green Bay Packers to three	FALSE	TRUE	1
1	48	In NFL, the red zone refers to the opponent team's end zone	FALSE	TRUE	1
1	49	In NFL, the first team to ever win the Super Bowl was the Green	FALSE	TRUE	2
1	50	In NFL, centres and linebackers are usually numbered between	FALSE	TRUE	1
1	51	Voluntary muscles are controlled by the cerebrum	FALSE	TRUE	2
1	52	Selective breeding results in a smaller gene pool	FALSE	TRUE	2
1	53	Suffocation is the main cause of death from avalanches	FALSE	TRUE	2
1	54	Hormones are transported around the body via the peripheral	FALSE	TRUE	1
1	55	Increased pressure speeds up the rate of chemical reactions in	FALSE	TRUE	1
1	56	Transition elements are commonly used as catalysts in chemical	FALSE	TRUE	2
1	57	The maximum number of electrons in the first electron shell is 2	FALSE	TRUE	2
1	58	Plant cells are easier to clone than animal cells	FALSE	TRUE	2
1	59	Secondary industries dominate the market in emerging	FALSE	TRUE	1
1	60	Scurvy and anemia are diseases not caused by bacteria or	FALSE	TRUE	2
1	61	Sedimentary rock typically form as a result of extremely high	FALSE	TRUE	1
1	62	Photosynthesis is an example of an endothermic reaction	FALSE	TRUE	2
1	63	Organisms working together to gain an advantage in	FALSE	TRUE	1
1	64	A catalyst has no effect on the equilibrium of a chemical system	FALSE	TRUE	2
1	65	Combustion reactions are typically irreversible	FALSE	TRUE	2

1	66	Epinephrine is a hormone prescribed to treat diabetes	FALSE	TRUE	1
1	67	A country's carbon footprint measures their impact on the	FALSE	TRUE	2
1	68	When contours on a map are very close together, it indicates a	FALSE	TRUE	2
1	69	Earthquakes and volcanoes typically occur at the boundaries of	FALSE	TRUE	2
1	70	When international aid is given directly from one country to	FALSE	TRUE	1
1	71	One disadvantage of biofuels is that they are carbon neutral	FALSE	TRUE	1
1	72	Deforestation is considered one of the main contributors to the	FALSE	TRUE	2
1	73	Sensory neurons are responsible for sending messages from the	FALSE	TRUE	1
1	74	The core of the earth is the hotter than its outer layers	FALSE	TRUE	2
1	75	Skin and bones are two organs that form part of the auxiliary	FALSE	TRUE	1
1	76	Some vaccines contain the live pathogen that has been treated	FALSE	TRUE	2
1	77	Elements in the standard periodic table are arranged in terms of	FALSE	TRUE	1
1	78	Bacteria are the primary source of energy for most food chains	FALSE	TRUE	1
1	79	The mass of an element equals the number of protons and	FALSE	TRUE	1
1	80	As the temperature increases, the solubility of gasses increases	FALSE	TRUE	1
1	81	There are four covalent bonds involved in a methane molecule	FALSE	TRUE	2
1	82	Electrolysis is the process of splitting substances up using heat	FALSE	TRUE	1
1	83	Random mutations in DNA are a common cause of mass	FALSE	TRUE	1
1	84	Most of the Nitrogen that plants use are absorbed through their	FALSE	TRUE	2
1	85	The amount of water in the body is regulated primarily via	FALSE	TRUE	1
1	86	Hurricanes usually only form if the sea temperature is less than	FALSE	TRUE	1
1	87	Isotopes have the same number of protons, but different	FALSE	TRUE	2
1	88	A substance with a pH of 8 is a strong acid	FALSE	TRUE	1
1	89	Pressure directly affects the solubility of gases in water	FALSE	TRUE	2
1	90	An alkali is a substance that produces hydrogen ions when	FALSE	TRUE	1
1	91	Hydrocarbates are essential to a balanced human diet	FALSE	TRUE	1
1	92	Selective breeding results in an increased risk of diseases in the	FALSE	TRUE	2
1	93	The main impact of deforestation is the loss of habitat for	FALSE	TRUE	2
1	94	A subsistence farmer is a farmer who focuses on raising animals	FALSE	TRUE	2
1	95	Antibiotics are ineffective against viruses	FALSE	TRUE	2
1	96	During photosynthesis, chemical energy is converted into light	FALSE	TRUE	1
1	97	Decane splitting into octane and ethene is an example of a	FALSE	TRUE	1
1	98	Darwin's theory was not widely accepted when it was first	FALSE	TRUE	2
1	99	Voluntary muscle messages are processed in the medulla of the	FALSE	TRUE	1

1	100	Metallic bonds are typically stronger than covalent bonds	FALSE	TRUE	1
2	1	In the 2018 NFL draft, Mark Andrews was drafted by the	FALSE	TRUE	1
2	2	In the 2018 NFL draft, Alabama was one of the schools with no	FALSE	TRUE	1
2	3	In the 2018 NFL draft, the New York Giants were the only team	FALSE	TRUE	1
2	4	In the 2018 NFL draft, Nathan Shepard was drafted by the New	FALSE	TRUE	2
2	5	In the 2018 NFL draft, the Baltimore Ravens were the team to	FALSE	TRUE	2
2	6	In the 2018 NFL draft, the Los Angeles Rams drafted Jordan	FALSE	TRUE	1
2	7	In the 2018 NFL draft, Terrell Edmunds was drafted by the	FALSE	TRUE	2
2	8	In the 2018 NFL draft, Baker Mayfield was drafted in the	FALSE	TRUE	1
2	9	In the 2018 NFL draft, the defending Super Bowl champions	FALSE	TRUE	2
2	10	In the 2018 NFL draft, the Washington Redskins were the team	FALSE	TRUE	2
2	11	In the 2017 NFL draft, the New England Patriots had only two	FALSE	TRUE	1
2	12	In the 2017 NFL draft, more than two players were selected	FALSE	TRUE	2
2	13	In the 2017 NFL draft, the Big Ten was one of the athletic	FALSE	TRUE	1
2	14	In the 2017 NFL draft, more than two players were drafted from	FALSE	TRUE	2
2	15	In the 2017 NFL draft, more than four players were drafted from	FALSE	TRUE	1
2	16	In the 2017 NFL draft, the Chicago Bears drafted Adam Shaheen	FALSE	TRUE	2
2	17	In the 2017 NFL draft, the New York Giants were penalized in	FALSE	TRUE	2
2	18	In the 2017 NFL draft, the Minnesota Vikings had less than 12	FALSE	TRUE	1
2	19	In the 2017 NFL draft, the four special teams players drafted	FALSE	TRUE	2
2	20	In the 2017 NFL draft, the Denver Broncos drafted Trey Quinn as	FALSE	TRUE	1
2	21	In the 2016 NFL draft, the Tennessee Titans were supposed to	FALSE	TRUE	1
2	22	In the 2016 NFL draft, Carson Wentz was the only player to be	FALSE	TRUE	1
2	23	In the 2016 NFL draft, wide receiver Mortiz Bohringer was	FALSE	TRUE	1
2	24	In the 2016 NFL draft, kicker Roberto Aguayo was drafted by the	FALSE	TRUE	2
2	25	In the 2016 NFL draft, the 2015 Heisman Trophy winner, Derrick	FALSE	TRUE	1
2	26	In the 2016 NFL draft, Ohio State University had the most	FALSE	TRUE	2
2	27	In the 2016 NFL draft, more than two teams were forced to	FALSE	TRUE	2
2	28	In the 2016 NFL draft, Rico Gathers was drafted by the Oakland	FALSE	TRUE	1
2	29	In the 2016 NFL draft, David Onyemata was drafted by the New	FALSE	TRUE	2
2	30	In the 2016 NFL draft, the Tennessee Titans drafted Kalan Reed	FALSE	TRUE	2
2	31	In NFL rules, a player performing a chop block will be penalized	FALSE	TRUE	1
2	32	In NFL rules, a player who wears illegal equipment is to be	FALSE	TRUE	1
2	33	In NFL rules, a delay of game penalty at the start of either half is	FALSE	TRUE	1

2	34	In NFL rules, the penalty for attempting to use more than 3	FALSE	TRUE	2
2	35	In NFL rules, when holding is committed by the defense, the	FALSE	TRUE	2
2	36	In NFL rules, the penalty for the first onside kick going out of	FALSE	TRUE	2
2	37	In NFL rules, pass interference and neutral zone infractions are	FALSE	TRUE	2
2	38	In NFL rules, the penalty for leaping is 5 yards when it is	FALSE	TRUE	1
2	39	In NFL rules, when an intentional grounding penalty is called at	FALSE	TRUE	2
2	40	In NFL rules, the penalty for any action which delays the next	FALSE	TRUE	1
2	41	In NFL, a sack is when a quarterback or other player on offense	FALSE	TRUE	1
2	42	In NFL, the quarterback is permitted to change the play if he	FALSE	TRUE	2
2	43	In NFL, a "three and out" refers to when three wide receivers set	FALSE	TRUE	1
2	44	In NFL, there 7 rounds in the draft, with the worst-performing NFL team from the previous year choosing first, and the best-	FALSE	TRUE	2
2	45	In NFL, a "Hail Mary" is a play in which the receivers are all sent	FALSE	TRUE	2
2	46	In NFL, a "two-point conversion" is a play a team attempts instead of kicking a one-point conversion immediately after it	FALSE	TRUE	2
2	47	In NFL, Vince Lombardi led the Green Bay Packers to three	FALSE	TRUE	1
2	48	In NFL, the red zone refers to the opponent team's end zone	FALSE	TRUE	1
2	49	In NFL, the first team to ever win the Super Bowl was the Green	FALSE	TRUE	2
2	50	In NFL, centres and linebackers are usually numbered between	FALSE	TRUE	1
2	51	Voluntary muscles are controlled by the cerebrum	FALSE	TRUE	2
2	52	Selective breeding results in a smaller gene pool	FALSE	TRUE	2
2	53	Suffocation is the main cause of death from avalanches	FALSE	TRUE	2
2	54	Hormones are transported around the body via the peripheral	FALSE	TRUE	1
2	55	Increased pressure speeds up the rate of chemical reactions in	FALSE	TRUE	1
2	56	Transition elements are commonly used as catalysts in chemical	FALSE	TRUE	2
2	57	The maximum number of electrons in the first electron shell is 2	FALSE	TRUE	2
2	58	Plant cells are easier to clone than animal cells	FALSE	TRUE	2
2	59	Secondary industries dominate the market in emerging	FALSE	TRUE	1
2	60	Scurvy and anemia are diseases not caused by bacteria or	FALSE	TRUE	2
2	61	Sedimentary rock typically form as a result of extremely high	FALSE	TRUE	1
2	62	Photosynthesis is an example of an endothermic reaction	FALSE	TRUE	2
2	63	Organisms working together to gain an advantage in	FALSE	TRUE	1
2	64	A catalyst has no effect on the equilibrium of a chemical system	FALSE	TRUE	2
2	65	Combustion reactions are typically irreversible	FALSE	TRUE	2

2	66	Epinephrine is a hormone prescribed to treat diabetes	FALSE	TRUE	1
2	67	A country's carbon footprint measures their impact on the	FALSE	TRUE	2
2	68	When contours on a map are very close together, it indicates a	FALSE	TRUE	2
2	69	Earthquakes and volcanoes typically occur at the boundaries of	FALSE	TRUE	2
2	70	When international aid is given directly from one country to	FALSE	TRUE	1
2	71	One disadvantage of biofuels is that they are carbon neutral	FALSE	TRUE	1
2	72	Deforestation is considered one of the main contributors to the	FALSE	TRUE	2
2	73	Sensory neurons are responsible for sending messages from the	FALSE	TRUE	1
2	74	The core of the earth is the hotter than its outer layers	FALSE	TRUE	2
2	75	Skin and bones are two organs that form part of the auxiliary	FALSE	TRUE	1
2	76	Some vaccines contain the live pathogen that has been treated	FALSE	TRUE	2
2	77	Elements in the standard periodic table are arranged in terms of	FALSE	TRUE	1
2	78	Bacteria are the primary source of energy for most food chains	FALSE	TRUE	1
2	79	The mass of an element equals the number of protons and	FALSE	TRUE	1
2	80	As the temperature increases, the solubility of gasses increases	FALSE	TRUE	1
2	81	There are four covalent bonds involved in a methane molecule	FALSE	TRUE	2
2	82	Electrolysis is the process of splitting substances up using heat	FALSE	TRUE	1
2	83	Random mutations in DNA are a common cause of mass	FALSE	TRUE	1
2	84	Most of the Nitrogen that plants use are absorbed through their	FALSE	TRUE	2
2	85	The amount of water in the body is regulated primarily via	FALSE	TRUE	1
2	86	Hurricanes usually only form if the sea temperature is less than	FALSE	TRUE	1
2	87	Isotopes have the same number of protons, but different	FALSE	TRUE	2
2	88	A substance with a pH of 8 is a strong acid	FALSE	TRUE	1
2	89	Pressure directly affects the solubility of gases in water	FALSE	TRUE	2
2	90	An alkali is a substance that produces hydrogen ions when	FALSE	TRUE	1
2	91	Hydrocarbates are essential to a balanced human diet	FALSE	TRUE	1
2	92	Selective breeding results in an increased risk of diseases in the	FALSE	TRUE	2
2	93	The main impact of deforestation is the loss of habitat for	FALSE	TRUE	2
2	94	A subsistence farmer is a farmer who focuses on raising animals	FALSE	TRUE	2
2	95	Antibiotics are ineffective against viruses	FALSE	TRUE	2
2	96	During photosynthesis, chemical energy is converted into light	FALSE	TRUE	1
2	97	Decane splitting into octane and ethene is an example of a	FALSE	TRUE	1
2	98	Darwin's theory was not widely accepted when it was first	FALSE	TRUE	2
2	99	Voluntary muscle messages are processed in the medulla of the	FALSE	TRUE	1

2	100	Metallic bonds are typically stronger than covalent bonds	FALSE	TRUE	1
2	101	Lee's workmate fails to deliver an important piece of information on time, causing Lee to fall behind schedule also.	Work harder to compensate	Explain the urgency of the situation to the	2
2	102	Rhea has left her job to be a full-time mother, which she loves, but she misses the company and companionship of her workmates. What action would be the most effective for Rhea?	Try to see her old workmates socially, inviting them out	Join a playgroup or social group of new mothers	2
2	103	Pete has specific skills that his workmates do not and he feels that his workload is higher because of it. What action would be	Speak to his workmates about this	Speak to his boss about this	2
2	104	Mario is showing Min, a new employee, how the system works. Mario's boss walks by and announces Mario is wrong about several points, as changes have been made. Mario gets on well	Learn the new changes	Make a joke to Min, explaining he didn't know about the changes	1
2	105	Wai-Hin and Connie have shared an office for years but Wai-Hin gets a new job and Connie loses contact with her. What action would be the most effective for Connie?	Contact Wai-Hin and arrange to catch up but also make friends with her replacement	Spend time getting to know the other people in the office, and strike up new friendships	1
2	106	Martina is accepted for a highly sought after contract, but has to fly to the location. Martina has a phobia of flying. What action	See a doctor about this	Find alternative travel arrangements	1
2	107	Manual is only a few years from retirement when he finds out his position will no longer exist, although he will still have a job with a less prestigious role. What action would be the most	Carefully consider his options and discuss it with his family	Talk to his boss or the management about it	1
2	108	Alan helps Trudy, a peer he works with occasionally, with a difficult task. Trudy complains that Alan's work isn't very good,	Apologize to Trudy	Diffuse the argument by asking for advice	2
2	109	Surbhi starts a new job where he doesn't know anyone and finds that no one is particularly friendly. What action would be the most effective for Surbhi?	Concentrate on doing his work well at the new job	Make an effort to talk to people and be friendly himself	2
2	110	Darla is nervous about presenting her work to a group of seniors who might not understand it, as they don't know much about her area. What action would be the most effective for Darla?	Work on her presentation, simplifying the explanations	Practice presenting to laypeople such as friends or family	1
2	111	Andre moves away from the city his friends and family are in. He finds his friends make less effort to keep in contact than he thought they would. What action would be the most effective	Try to adjust to life in the new city by joining clubs and activities there	He should make the effort to contact them, but also try to meet	2
2	112	Helga's team has been performing very well. They receive poor-quality work from another team that they must incorporate into	Tell the project manager about the situation	Tell the other team they must re-do their work	1

2	113	Clayton has been overseas for a long time and returns to visit his family. So much has changed that Clayton feels left out.	Spend time listening and getting involved again	Tell his family he feels left out	1
2	114	Katerina takes a long time to set the DVD timer. With the family watching, her sister says ?You idiot, you're doing it all wrong,	Ignore her sister and keep at the task	Get her sister to help or to do it	2
2	115	Benjiro's parents are in their late 80s and living interstate in a house by themselves. He is worried that they need some help but they angrily deny it any time he brings up the subject. What	Keep telling his parents his concerns, stressing their importance	Visit frequently and get others to check on them	2
2	116	Max prides himself on his work being of the highest quality. On a joint project, other people do a lousy job, assuming that Max will fix their mistakes. What action would be the most effective	Tell the project manager about the situation	Confront the others, and tell them they must fix their mistakes	2
2	117	Daniel has been accepted for a prestigious position in a different country from his family, who he is close to. He and his wife decide it is worth relocating. What action would be the most	Set up a system for staying in touch, like weekly phone calls or emails	Think about the great opportunities this change offers	1
2	118	A junior employee making routine adjustments to some of Teo's equipment accuses Teo of causing the equipment malfunction.	Explain that malfunctions were not his fault	Ignore the accusation, it is not important	1
2	119	Mei Ling answers the phone and hears that close relatives are in hospital critically ill. What action would be the most effective for Mei Ling?	Let herself cry and express emotion for as long as she feels like	Speak to other family to calm herself and find out what is happening, then visit the hospital	2
2	120	The woman who relieves Celia at the end of her shift is twenty minutes late without excuse or apology. What action would be	Ask for an explanation of her lateness	Tell her that this is unacceptable	1
2	121	Upon entering full-time study, Vincent cannot afford the time or money he used to spend on water-polo training, which he was quite good at. Although he enjoys full-time study, he misses	Find out about sporting scholarships or bursaries	See if there is a local league or a less expensive and less time-consuming	2
2	122	Evan's housemate cooked food late at night and left a huge mess in the kitchen that Evan discovered at breakfast. What	Ask his housemate that this not happen again	Tell his housemate to clean up the mess	1
2	123	Greg has just gone back to university after a lapse of several years. He is surrounded by younger students who seem very	Talk to others in his situation	Study hard and attend all lectures	1
2	124	Gloria's housemates never buy essential non-food items when they are running low, relying on Gloria to buy them, which she resents. They know each other reasonably well, but have not yet	Introduce a new system for grocery shopping and sharing costs	Tell her housemates she has a problem with this	2

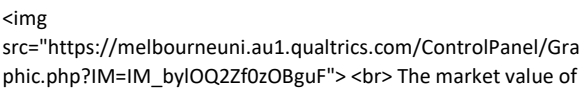
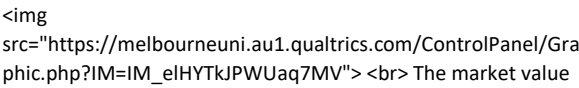
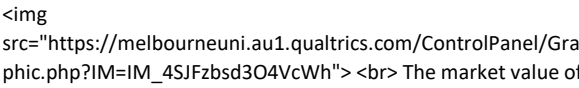
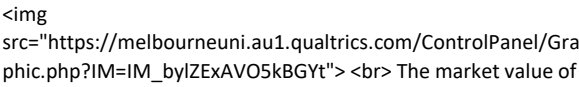
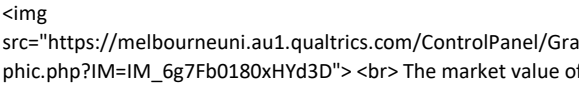
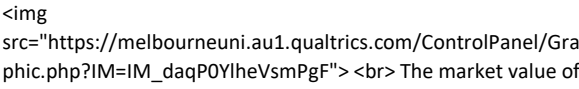
2	125	Shona has not spoken to her nephew for months, whereas when he was younger they were very close. She rings him but he can only talk for five minutes. What action would be the most effective for Shona?	Make plans to drop by and visit him in person and have a good chat	Understand that relationships change, but keep calling him from time to time	2
2	126	Moshe finds out that some members of his social sports team have been saying that he is not a very good player. What action would be the most effective for Moshe?	Do some extra training to try and improve	Although he may be bad at sport, remember he is good at other things	1
2	127	Joel has always dealt with one particular client but on a very complex job his boss gives the task to a co-worker instead. Joel wonders whether his boss thinks he can't handle the important	Ask his boss why the co-worker was given the job	Do good work so that he will be given the complex tasks in future	1
2	128	Hasina is overseas when she finds out that her father has passed away from an illness he has had for years. What action would be the most effective for Hasina?	Contact her close relatives for information and support	Think deeply about the more profound meaning of this loss	1
2	129	Mina and her sister-in-law normally get along quite well, and the sister-in-law regularly baby-sits for her for a small fee. Lately she has also been cleaning away cobwebs, commenting on the	Tell her only to baby-sit, not to clean	Tell her sister-in-law these comments upset her	2
2	130	Billy is nervous about acting a scene when there are a lot of very experienced actors in the crowd. What action would be the	Believe in himself and know it will be fine	Use some acting techniques to calm his	2
2	131	Juno is fairly sure his company is going down and his job is under threat. It is a large company and nothing official has been said. What action would be the most effective for Juno?	Find out what is happening and discuss his concerns with his family	Think of these events as an opportunity for a new start	1
2	132	Mallory moves from a small company to a very large one, where there is little personal contact, which she misses. What action would be the most effective for Mallory?	Talk to her workmates, try to create social contacts and make friends	Concentrate on her outside-work friends and colleagues from previous	1
2	133	A demanding client takes up a lot of Jill's time and then asks to speak to Jill's boss about her performance. Although Jill's boss assures her that her performance is fine, Jill feels upset. What action would be the most effective for Jill?	Calm down by taking deep breaths or going for a short walk	Think that she has been successful in the past and this client being difficult is not her fault	2
2	134	Blair and Flynn usually go to a cafe after the working week and chat about what's going on in the company. After Blair's job is moved to a different section in the company, he stops coming to	Invite Blair again, maybe rescheduling for another time	Go to the cafe or socialize with other workers	1
2	135	Jerry has had several short-term jobs in the same industry, but is excited about starting a job in a different industry. His father casually remarks that he will probably last six months. What	Prove him wrong by working hard to succeed at the new job	Tell his father he is completely wrong	1

2	136	Michelle's friend Dara is moving overseas to live with her partner. They have been good friends for many years and Dara is unlikely to come back. What action would be the most	Spend time with other friends, keeping herself busy	Make sure she keeps in contact through email, phone or letter writing	2
2	137	Dorian needs to have some prostate surgery and is quite scared about the process. He has heard that it is quite painful. What action would be the most effective for Dorian?	Keep busy in the meantime so he doesn't think about the impending	Talk to his doctor about what will happen	2
2	138	Hannah's access to essential resources has been delayed and her work is way behind schedule. Her progress report makes no mention of the lack of resources. What action would be the	Explain the lack of resources to her boss or to management	Document the lack of resources in her progress report	2
2	139	Jill is given an official warning for entering a restricted area. She was never informed that the area was restricted and will lose her job if she gets two more warnings, which she thinks is unfair. What action would be the most effective for Jill?	Accept the warning and be careful not to go in restricted areas from now on	Explain that she didn't know it was restricted	2
2	140	Alana has been acting in a high-ranking role for several months. A decision is made that only long-term employees can now act	Quit that position	Ask management if an exception can be made	2
2	141	Reece's friend points out that her young children seem to be developing more quickly than Reece's. Reece sees that this is true. What action would be the most effective for Reece?	Talk to a doctor about what the normal rates of development are	Realize that children develop at different rates	1
2	142	Jumah has been working at a new job part-time while he studies. His shift times for the week are changed at the last minute, without consulting him. What action would be the most	Tell the manager in charge of shifts that he is not happy about it	Find out if there is some reasonable explanation for the shift changes	2
2	143	Jacob is having a large family gathering to celebrate him moving into his new home. He wants the day to go smoothly and is a little nervous about it. What action would be the most effective	Prepare ahead of time so he has everything he needs available	Accept that things aren't going to be perfect but the family will	1
2	144	Julie hasn't seen Ka for ages and looks forward to their weekend trip away. However, Ka has changed a lot and Julie finds that she is no longer an interesting companion. What	Understand that people change, so move on, but remember the good times	Concentrate on her other, more rewarding friendships	1
2	145	Song finds out that a friend of hers has borrowed money from others to pay urgent bills, but has in fact used the money for	Anger	Contempt	2
2	146	Edna's workmate organizes a goodbye party for Edna, who is	Gratitude	Surprise	1
2	147	Kevin has been working at his current job for a few years. Out of the blue, he finds that he will receive a promotion. Kevin is most	Joy	Pride	1
2	148	Garry's small business is attracting less and less clients and he can't tell why. There doesn't seem to be anything he can do to	Sad	Distressed	2

2	149	The new manager at Enid's work changes everyone's hours to a less flexible work pattern, leaving no room for discussion. Enid is	Dislike	Rage	1
2	150	Rashid needs to meet a quota before his performance review. There is only a small chance that he will be able to do so and	Hopeful	Scared	2
3	1	phic.php?IM=IM_8qtY35T40bAycjH"> The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	7	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	8	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	9	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1

3	10	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	11	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	12	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	13	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	14	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	15	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1

3	16	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	17	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	18	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	19	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	20	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
3	21	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2

3	22	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	23	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	24	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	25	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	26	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	27	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2

3	28	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	29	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	30	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	31	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	32	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	33	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2

3	34	                                               
---	----	--

3	40	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
3	41	Voluntary muscles are controlled by the cerebrum	FALSE	TRUE	2
3	42	Hormones are transported around the body via the peripheral nervous system	FALSE	TRUE	1
3	43	Increased pressure speeds up the rate of chemical reactions in both gases and liquids	FALSE	TRUE	1
3	44	Transition elements are commonly used as catalysts in chemical reactions	FALSE	TRUE	2
3	45	The maximum number of electrons in the first electron shell is 2	FALSE	TRUE	2
3	46	Plant cells are easier to clone than animal cells	FALSE	TRUE	2
3	47	Secondary industries dominate the market in emerging economies	FALSE	TRUE	1
3	48	Scurvy and anemia are diseases not caused by bacteria or viruses	FALSE	TRUE	2
3	49	Sedimentary rock typically form as a result of extremely high pressure or heat	FALSE	TRUE	1
3	50	Photosynthesis is an example of an endothermic reaction	FALSE	TRUE	2
3	51	Organisms working together to gain an advantage in competition is an example of parasitism	FALSE	TRUE	1
3	52	A catalyst has no effect on the equilibrium of a chemical system	FALSE	TRUE	2
3	53	Epinephrine is a hormone prescribed to treat diabetes	FALSE	TRUE	1
3	54	When contours on a map are very close together, it indicates a steep slope	FALSE	TRUE	2
3	55	When international aid is given directly from one country to another, it is called multilateral aid	FALSE	TRUE	1
3	56	One disadvantage of biofuels is that they are carbon neutral	FALSE	TRUE	1

3	57	Deforestation is considered one of the main contributors to the greenhouse effect	FALSE	TRUE	2
3	58	Sensory neurons are responsible for sending messages from the brain to different glands in the body	FALSE	TRUE	1
3	59	Skin and bones are two organs that form part of the auxiliary nervous system	FALSE	TRUE	1
3	60	Some vaccines contain the live pathogen that has been treated to make it harmless	FALSE	TRUE	2
3	61	Elements in the standard periodic table are arranged in terms of atomic mass	FALSE	TRUE	1
3	62	Bacteria are the primary source of energy for most food chains	FALSE	TRUE	1
3	63	The mass of an element equals the number of protons and electrons in one atom	FALSE	TRUE	1
3	64	As the temperature increases, the solubility of gasses increases	FALSE	TRUE	1
3	65	There are four covalent bonds involved in a methane molecule	FALSE	TRUE	2
3	66	Electrolysis is the process of splitting substances up using heat energy	FALSE	TRUE	1
3	67	Random mutations in DNA are a common cause of mass extinction in a species	FALSE	TRUE	1
3	68	Most of the Nitrogen that plants use are absorbed through their roots	FALSE	TRUE	2
3	69	The amount of water in the body is regulated primarily via thyroid gland	FALSE	TRUE	1
3	70	Hurricanes usually only form if the sea temperature is less than 80 degrees Fahrenheit	FALSE	TRUE	1
3	71	A substance with a pH of 8 is a strong acid	FALSE	TRUE	1
3	72	An alkali is a substance that produces hydrogen ions when added to water	FALSE	TRUE	1
3	73	Hydrocarbates are essential to a balanced human diet	FALSE	TRUE	1
3	74	Selective breeding results in an increased risk of diseases in the population	FALSE	TRUE	2

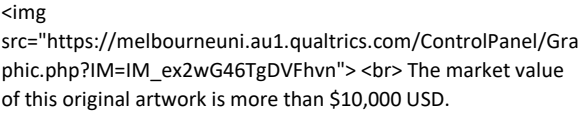
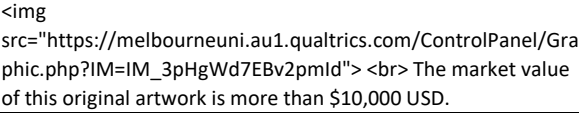
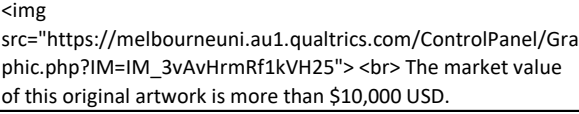
3	75	Antibiotics are ineffective against viruses	FALSE	TRUE	2
3	76	During photosynthesis, chemical energy is converted into light energy	FALSE	TRUE	1
3	77	Decane splitting into octane and ethene is an example of a polymerization reaction	FALSE	TRUE	1
3	78	Darwin's theory was not widely accepted when it was first published in the late 19th century	FALSE	TRUE	2
3	79	Voluntary muscle messages are processed in the medulla of the brain	FALSE	TRUE	1
3	80	Metallic bonds are typically stronger than covalent bonds	FALSE	TRUE	1
3	81	Lee's workmate fails to deliver an important piece of information on time, causing Lee to fall behind schedule also. What action would be the most effective for Lee?	Work harder to compensate	Explain the urgency of the situation to the workmate	2
3	82	Rhea has left her job to be a full-time mother, which she loves, but she misses the company and companionship of her workmates. What action would be the most effective for Rhea?	Try to see her old workmates socially, inviting them out	Join a playgroup or social group of new mothers	2
3	83	Pete has specific skills that his workmates do not and he feels that his workload is higher because of it. What action would be the most effective for Pete?	Speak to his workmates about this	Speak to his boss about this	2
3	84	Mario is showing Min, a new employee, how the system works. Mario's boss walks by and announces Mario is wrong about several points, as changes have been made. Mario gets on well with his boss, although they don't normally have much to do with each other. What action would be the most effective for Mario?	Learn the new changes	Make a joke to Min, explaining he didn't know about the changes	1
3	85	Wai-Hin and Connie have shared an office for years but Wai-Hin gets a new job and Connie loses contact with her. What action would be the most effective for Connie?	Contact Wai-Hin and arrange to catch up but also make friends with her replacement	Spend time getting to know the other people in the office, and strike up new friendships	1
3	86	Martina is accepted for a highly sought after contract, but has to fly to the location. Martina has a phobia of flying. What action would be the most effective for Martina?	See a doctor about this	Find alternative travel arrangements	1

3	87	Manual is only a few years from retirement when he finds out his position will no longer exist, although he will still have a job with a less prestigious role. What action would be the most effective for Manual?	Carefully consider his options and discuss it with his family	Talk to his boss or the management about it	1
3	88	Alan helps Trudy, a peer he works with occasionally, with a difficult task. Trudy complains that Alan's work isn't very good, and Alan responds that Trudy should be grateful he is doing her a favor. They argue. What action would be the most effective for Alan?	Apologize to Trudy	Diffuse the argument by asking for advice	2
3	89	Surbhi starts a new job where he doesn't know anyone and finds that no one is particularly friendly. What action would be the most effective for Surbhi?	Concentrate on doing his work well at the new job	Make an effort to talk to people and be friendly himself	2
3	90	Darla is nervous about presenting her work to a group of seniors who might not understand it, as they don't know much about her area. What action would be the most effective for Darla?	Work on her presentation, simplifying the explanations	Practice presenting to laypeople such as friends or family	1
3	91	Andre moves away from the city his friends and family are in. He finds his friends make less effort to keep in contact than he thought they would. What action would be the most effective for Andre?	Try to adjust to life in the new city by joining clubs and activities there	He should make the effort to contact them, but also try to meet people in his new city	2
3	92	Helga's team has been performing very well. They receive poor-quality work from another team that they must incorporate into their own project. What action would be the most effective for Helga?	Tell the project manager about the situation	Tell the other team they must re-do their work	1
3	93	Clayton has been overseas for a long time and returns to visit his family. So much has changed that Clayton feels left out. What action would be the most effective for Clayton?	Spend time listening and getting involved again	Tell his family he feels left out	1
3	94	Max prides himself on his work being of the highest quality. On a joint project, other people do a lousy job, assuming that Max will fix their mistakes. What action would be the most effective for Max?	Tell the project manager about the situation	Confront the others, and tell them they must fix their mistakes	2

3	95	Daniel has been accepted for a prestigious position in a different country from his family, who he is close to. He and his wife decide it is worth relocating. What action would be the most effective for Daniel?	Set up a system for staying in touch, like weekly phone calls or emails	Think about the great opportunities this change offers	1
3	96	The woman who relieves Celia at the end of her shift is twenty minutes late without excuse or apology. What action would be the most effective for Celia?	Ask for an explanation of her lateness	Tell her that this is unacceptable	1
3	97	Upon entering full-time study, Vincent cannot afford the time or money he used to spend on water-polo training, which he was quite good at. Although he enjoys full-time study, he misses training. What action would be the most effective for Vincent?	Find out about sporting scholarships or bursaries	See if there is a local league or a less expensive and less time-consuming sport	2
3	98	Evan's housemate cooked food late at night and left a huge mess in the kitchen that Evan discovered at breakfast. What action would be the most effective for Evan?	Ask his housemate that this not happen again	Tell his housemate to clean up the mess	1
3	99	Greg has just gone back to university after a lapse of several years. He is surrounded by younger students who seem very confident about their ability and he is unsure whether he can compete with them. What action would be the most effective for Greg?	Talk to others in his situation	Study hard and attend all lectures	1
3	100	Gloria's housemates never buy essential non-food items when they are running low, relying on Gloria to buy them, which she resents. They know each other reasonably well, but have not yet discussed financial issues. What action would be the most effective for Gloria?	Introduce a new system for grocery shopping and sharing costs	Tell her housemates she has a problem with this	2
3	101	Shona has not spoken to her nephew for months, whereas when he was younger they were very close. She rings him but he can only talk for five minutes. What action would be the most effective for Shona?	Make plans to drop by and visit him in person and have a good chat	Understand that relationships change, but keep calling him from time to time	2

3	102	Moshe finds out that some members of his social sports team have been saying that he is not a very good player. What action would be the most effective for Moshe?	Do some extra training to try and improve	Although he may be bad at sport, remember he is good at other things	1
3	103	Joel has always dealt with one particular client but on a very complex job his boss gives the task to a co-worker instead. Joel wonders whether his boss thinks he can't handle the important jobs. What action would be the most effective for Joel?	Ask his boss why the co-worker was given the job	Do good work so that he will be given the complex tasks in future	1
3	104	Hasina is overseas when she finds out that her father has passed away from an illness he has had for years. What action would be the most effective for Hasina?	Contact her close relatives for information and support	Think deeply about the more profound meaning of this loss	1
3	105	Mina and her sister-in-law normally get along quite well, and the sister-in-law regularly baby-sits for her for a small fee. Lately she has also been cleaning away cobwebs, commenting on the mess, which Mina finds insulting. What action would be the most effective for Mina?	Tell her only to baby-sit, not to clean	Tell her sister-in-law these comments upset her	2
3	106	Billy is nervous about acting a scene when there are a lot of very experienced actors in the crowd. What action would be the most effective for Billy?	Believe in himself and know it will be fine	Use some acting techniques to calm his nerves	2
3	107	Juno is fairly sure his company is going down and his job is under threat. It is a large company and nothing official has been said. What action would be the most effective for Juno?	Find out what is happening and discuss his concerns with his family	Think of these events as an opportunity for a new start	1
3	108	A demanding client takes up a lot of Jill's time and then asks to speak to Jill's boss about her performance. Although Jill's boss assures her that her performance is fine, Jill feels upset. What action would be the most effective for Jill?	Calm down by taking deep breaths or going for a short walk	Think that she has been successful in the past and this client being difficult is not her fault	2

3	109	Blair and Flynn usually go to a cafe after the working week and chat about what's going on in the company. After Blair's job is moved to a different section in the company, he stops coming to the cafe. Flynn misses these Friday talks. What action would be the most effective for Flynn?	Invite Blair again, maybe rescheduling for another time	Go to the cafe or socialize with other workers	1
3	110	Michelle's friend Dara is moving overseas to live with her partner. They have been good friends for many years and Dara is unlikely to come back. What action would be the most effective for Michelle?	Spend time with other friends, keeping herself busy	Make sure she keeps in contact through email, phone or letter writing	2
3	111	Hannah's access to essential resources has been delayed and her work is way behind schedule. Her progress report makes no mention of the lack of resources. What action would be the most effective for Hannah?	Explain the lack of resources to her boss or to management	Document the lack of resources in her progress report	2
3	112	Jill is given an official warning for entering a restricted area. She was never informed that the area was restricted and will lose her job if she gets two more warnings, which she thinks is unfair. What action would be the most effective for Jill?	Accept the warning and be careful not to go in restricted areas from now on	Explain that she didn't know it was restricted	2
3	113	Alana has been acting in a high-ranking role for several months. A decision is made that only long-term employees can now act in these roles, and Alana has not been with the company long enough to do so. What action would be the most effective for Alana?	Quit that position	Ask management if an exception can be made	2
3	114	Reece's friend points out that her young children seem to be developing more quickly than Reece's. Reece sees that this is true. What action would be the most effective for Reece?	Talk to a doctor about what the normal rates of development are	Realize that children develop at different rates	1
3	115	Jumah has been working at a new job part-time while he studies. His shift times for the week are changed at the last minute, without consulting him. What action would be the most effective for Jumah?	Tell the manager in charge of shifts that he is not happy about it	Find out if there is some reasonable explanation for the shift changes	2

3	116	Jacob is having a large family gathering to celebrate him moving into his new home. He wants the day to go smoothly and is a little nervous about it. What action would be the most effective for Jacob?	Prepare ahead of time so he has everything he needs available	Accept that things aren't going to be perfect but the family will understand	1
3	117	Song finds out that a friend of hers has borrowed money from others to pay urgent bills, but has in fact used the money for less serious purposes. Song is most likely to feel?	Anger	Contempt	2
3	118	Edna's workmate organizes a goodbye party for Edna, who is going on holidays. Edna is most likely to feel?	Gratitude	Surprise	1
3	119	Kevin has been working at his current job for a few years. Out of the blue, he finds that he will receive a promotion. Kevin is most likely to feel?	Joy	Pride	1
3	120	Rashid needs to meet a quota before his performance review. There is only a small chance that he will be able to do so and there isn't much he can do to improve the outcome. Rashid is most likely to feel?	Hopeful	Scared	2
4	1	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	2	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	3	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1

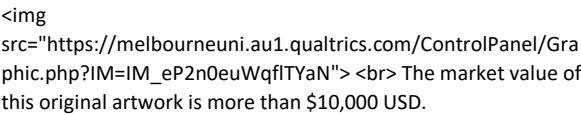
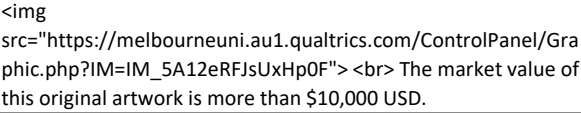
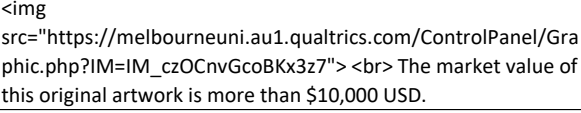
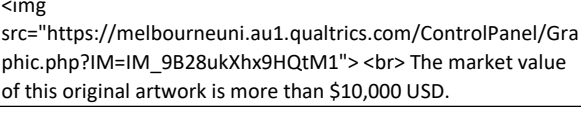
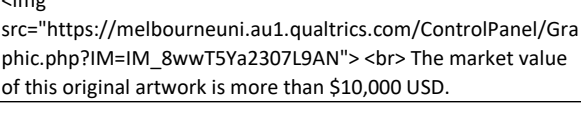
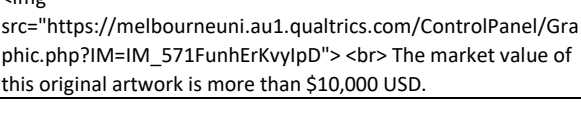
4	4	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	5	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	6	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	7	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	8	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	9	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1

4	10	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	11	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	12	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	13	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	14	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	15	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1

4	16	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	17	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	18	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	19	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	20	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	1
4	21	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2

4	22	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	23	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	24	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	25	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	26	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	27	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2

4	28	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	29	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	30	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	31	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	32	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	33	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2

4	34		FALSE	TRUE	2
4	35		FALSE	TRUE	2
4	36		FALSE	TRUE	2
4	37		FALSE	TRUE	2
4	38		FALSE	TRUE	2
4	39		FALSE	TRUE	2

4	40	 The market value of this original artwork is more than \$10,000 USD.	FALSE	TRUE	2
4	41	Voluntary muscles are controlled by the cerebrum	FALSE	TRUE	2
4	42	Hormones are transported around the body via the peripheral nervous system	FALSE	TRUE	1
4	43	Increased pressure speeds up the rate of chemical reactions in both gases and liquids	FALSE	TRUE	1
4	44	Transition elements are commonly used as catalysts in chemical reactions	FALSE	TRUE	2
4	45	The maximum number of electrons in the first electron shell is 2	FALSE	TRUE	2
4	46	Plant cells are easier to clone than animal cells	FALSE	TRUE	2
4	47	Secondary industries dominate the market in emerging economies	FALSE	TRUE	1
4	48	Scurvy and anemia are diseases not caused by bacteria or viruses	FALSE	TRUE	2
4	49	Sedimentary rock typically form as a result of extremely high pressure or heat	FALSE	TRUE	1
4	50	Photosynthesis is an example of an endothermic reaction	FALSE	TRUE	2
4	51	Organisms working together to gain an advantage in competition is an example of parasitism	FALSE	TRUE	1
4	52	A catalyst has no effect on the equilibrium of a chemical system	FALSE	TRUE	2
4	53	Epinephrine is a hormone prescribed to treat diabetes	FALSE	TRUE	1
4	54	When contours on a map are very close together, it indicates a steep slope	FALSE	TRUE	2
4	55	When international aid is given directly from one country to another, it is called multilateral aid	FALSE	TRUE	1
4	56	One disadvantage of biofuels is that they are carbon neutral	FALSE	TRUE	1

4	57	Deforestation is considered one of the main contributors to the greenhouse effect	FALSE	TRUE	2
4	58	Sensory neurons are responsible for sending messages from the brain to different glands in the body	FALSE	TRUE	1
4	59	Skin and bones are two organs that form part of the auxiliary nervous system	FALSE	TRUE	1
4	60	Some vaccines contain the live pathogen that has been treated to make it harmless	FALSE	TRUE	2
4	61	Elements in the standard periodic table are arranged in terms of atomic mass	FALSE	TRUE	1
4	62	Bacteria are the primary source of energy for most food chains	FALSE	TRUE	1
4	63	The mass of an element equals the number of protons and electrons in one atom	FALSE	TRUE	1
4	64	As the temperature increases, the solubility of gasses increases	FALSE	TRUE	1
4	65	There are four covalent bonds involved in a methane molecule	FALSE	TRUE	2
4	66	Electrolysis is the process of splitting substances up using heat energy	FALSE	TRUE	1
4	67	Random mutations in DNA are a common cause of mass extinction in a species	FALSE	TRUE	1
4	68	Most of the Nitrogen that plants use are absorbed through their roots	FALSE	TRUE	2
4	69	The amount of water in the body is regulated primarily via thyroid gland	FALSE	TRUE	1
4	70	Hurricanes usually only form if the sea temperature is less than 80 degrees Fahrenheit	FALSE	TRUE	1
4	71	A substance with a pH of 8 is a strong acid	FALSE	TRUE	1
4	72	An alkali is a substance that produces hydrogen ions when added to water	FALSE	TRUE	1
4	73	Hydrocarbates are essential to a balanced human diet	FALSE	TRUE	1
4	74	Selective breeding results in an increased risk of diseases in the population	FALSE	TRUE	2

4	75	Antibiotics are ineffective against viruses	FALSE	TRUE	2
4	76	During photosynthesis, chemical energy is converted into light energy	FALSE	TRUE	1
4	77	Decane splitting into octane and ethene is an example of a polymerization reaction	FALSE	TRUE	1
4	78	Darwin's theory was not widely accepted when it was first published in the late 19th century	FALSE	TRUE	2
4	79	Voluntary muscle messages are processed in the medulla of the brain	FALSE	TRUE	1
4	80	Metallic bonds are typically stronger than covalent bonds	FALSE	TRUE	1
4	81	Lee's workmate fails to deliver an important piece of information on time, causing Lee to fall behind schedule also. What action would be the most effective for Lee?	Work harder to compensate	Explain the urgency of the situation to the workmate	2
4	82	Rhea has left her job to be a full-time mother, which she loves, but she misses the company and companionship of her workmates. What action would be the most effective for Rhea?	Try to see her old workmates socially, inviting them out	Join a playgroup or social group of new mothers	2
4	83	Pete has specific skills that his workmates do not and he feels that his workload is higher because of it. What action would be the most effective for Pete?	Speak to his workmates about this	Speak to his boss about this	2
4	84	Mario is showing Min, a new employee, how the system works. Mario's boss walks by and announces Mario is wrong about several points, as changes have been made. Mario gets on well with his boss, although they don't normally have much to do with each other. What action would be the most effective for Mario?	Learn the new changes	Make a joke to Min, explaining he didn't know about the changes	1
4	85	Wai-Hin and Connie have shared an office for years but Wai-Hin gets a new job and Connie loses contact with her. What action would be the most effective for Connie?	Contact Wai-Hin and arrange to catch up but also make friends with her replacement	Spend time getting to know the other people in the office, and strike up new friendships	1
4	86	Martina is accepted for a highly sought after contract, but has to fly to the location. Martina has a phobia of flying. What action would be the most effective for Martina?	See a doctor about this	Find alternative travel arrangements	1

4	87	Manual is only a few years from retirement when he finds out his position will no longer exist, although he will still have a job with a less prestigious role. What action would be the most effective for Manual?	Carefully consider his options and discuss it with his family	Talk to his boss or the management about it	1
4	88	Alan helps Trudy, a peer he works with occasionally, with a difficult task. Trudy complains that Alan's work isn't very good, and Alan responds that Trudy should be grateful he is doing her a favor. They argue. What action would be the most effective for Alan?	Apologize to Trudy	Diffuse the argument by asking for advice	2
4	89	Surbhi starts a new job where he doesn't know anyone and finds that no one is particularly friendly. What action would be the most effective for Surbhi?	Concentrate on doing his work well at the new job	Make an effort to talk to people and be friendly himself	2
4	90	Darla is nervous about presenting her work to a group of seniors who might not understand it, as they don't know much about her area. What action would be the most effective for Darla?	Work on her presentation, simplifying the explanations	Practice presenting to laypeople such as friends or family	1
4	91	Andre moves away from the city his friends and family are in. He finds his friends make less effort to keep in contact than he thought they would. What action would be the most effective for Andre?	Try to adjust to life in the new city by joining clubs and activities there	He should make the effort to contact them, but also try to meet people in his new city	2
4	92	Helga's team has been performing very well. They receive poor-quality work from another team that they must incorporate into their own project. What action would be the most effective for Helga?	Tell the project manager about the situation	Tell the other team they must re-do their work	1
4	93	Clayton has been overseas for a long time and returns to visit his family. So much has changed that Clayton feels left out. What action would be the most effective for Clayton?	Spend time listening and getting involved again	Tell his family he feels left out	1
4	94	Max prides himself on his work being of the highest quality. On a joint project, other people do a lousy job, assuming that Max will fix their mistakes. What action would be the most effective for Max?	Tell the project manager about the situation	Confront the others, and tell them they must fix their mistakes	2

4	95	Daniel has been accepted for a prestigious position in a different country from his family, who he is close to. He and his wife decide it is worth relocating. What action would be the most effective for Daniel?	Set up a system for staying in touch, like weekly phone calls or emails	Think about the great opportunities this change offers	1
4	96	The woman who relieves Celia at the end of her shift is twenty minutes late without excuse or apology. What action would be the most effective for Celia?	Ask for an explanation of her lateness	Tell her that this is unacceptable	1
4	97	Upon entering full-time study, Vincent cannot afford the time or money he used to spend on water-polo training, which he was quite good at. Although he enjoys full-time study, he misses training. What action would be the most effective for Vincent?	Find out about sporting scholarships or bursaries	See if there is a local league or a less expensive and less time-consuming sport	2
4	98	Evan's housemate cooked food late at night and left a huge mess in the kitchen that Evan discovered at breakfast. What action would be the most effective for Evan?	Ask his housemate that this not happen again	Tell his housemate to clean up the mess	1
4	99	Greg has just gone back to university after a lapse of several years. He is surrounded by younger students who seem very confident about their ability and he is unsure whether he can compete with them. What action would be the most effective for Greg?	Talk to others in his situation	Study hard and attend all lectures	1
4	100	Gloria's housemates never buy essential non-food items when they are running low, relying on Gloria to buy them, which she resents. They know each other reasonably well, but have not yet discussed financial issues. What action would be the most effective for Gloria?	Introduce a new system for grocery shopping and sharing costs	Tell her housemates she has a problem with this	2
4	101	Shona has not spoken to her nephew for months, whereas when he was younger they were very close. She rings him but he can only talk for five minutes. What action would be the most effective for Shona?	Make plans to drop by and visit him in person and have a good chat	Understand that relationships change, but keep calling him from time to time	2

4	102	Moshe finds out that some members of his social sports team have been saying that he is not a very good player. What action would be the most effective for Moshe?	Do some extra training to try and improve	Although he may be bad at sport, remember he is good at other things	1
4	103	Joel has always dealt with one particular client but on a very complex job his boss gives the task to a co-worker instead. Joel wonders whether his boss thinks he can't handle the important jobs. What action would be the most effective for Joel?	Ask his boss why the co-worker was given the job	Do good work so that he will be given the complex tasks in future	1
4	104	Hasina is overseas when she finds out that her father has passed away from an illness he has had for years. What action would be the most effective for Hasina?	Contact her close relatives for information and support	Think deeply about the more profound meaning of this loss	1
4	105	Mina and her sister-in-law normally get along quite well, and the sister-in-law regularly baby-sits for her for a small fee. Lately she has also been cleaning away cobwebs, commenting on the mess, which Mina finds insulting. What action would be the most effective for Mina?	Tell her only to baby-sit, not to clean	Tell her sister-in-law these comments upset her	2
4	106	Billy is nervous about acting a scene when there are a lot of very experienced actors in the crowd. What action would be the most effective for Billy?	Believe in himself and know it will be fine	Use some acting techniques to calm his nerves	2
4	107	Juno is fairly sure his company is going down and his job is under threat. It is a large company and nothing official has been said. What action would be the most effective for Juno?	Find out what is happening and discuss his concerns with his family	Think of these events as an opportunity for a new start	1
4	108	A demanding client takes up a lot of Jill's time and then asks to speak to Jill's boss about her performance. Although Jill's boss assures her that her performance is fine, Jill feels upset. What action would be the most effective for Jill?	Calm down by taking deep breaths or going for a short walk	Think that she has been successful in the past and this client being difficult is not her fault	2

4	109	Blair and Flynn usually go to a cafe after the working week and chat about what's going on in the company. After Blair's job is moved to a different section in the company, he stops coming to the cafe. Flynn misses these Friday talks. What action would be the most effective for Flynn?	Invite Blair again, maybe rescheduling for another time	Go to the cafe or socialize with other workers	1
4	110	Michelle's friend Dara is moving overseas to live with her partner. They have been good friends for many years and Dara is unlikely to come back. What action would be the most effective for Michelle?	Spend time with other friends, keeping herself busy	Make sure she keeps in contact through email, phone or letter writing	2
4	111	Hannah's access to essential resources has been delayed and her work is way behind schedule. Her progress report makes no mention of the lack of resources. What action would be the most effective for Hannah?	Explain the lack of resources to her boss or to management	Document the lack of resources in her progress report	2
4	112	Jill is given an official warning for entering a restricted area. She was never informed that the area was restricted and will lose her job if she gets two more warnings, which she thinks is unfair. What action would be the most effective for Jill?	Accept the warning and be careful not to go in restricted areas from now on	Explain that she didn't know it was restricted	2
4	113	Alana has been acting in a high-ranking role for several months. A decision is made that only long-term employees can now act in these roles, and Alana has not been with the company long enough to do so. What action would be the most effective for Alana?	Quit that position	Ask management if an exception can be made	2
4	114	Reece's friend points out that her young children seem to be developing more quickly than Reece's. Reece sees that this is true. What action would be the most effective for Reece?	Talk to a doctor about what the normal rates of development are	Realize that children develop at different rates	1
4	115	Jumah has been working at a new job part-time while he studies. His shift times for the week are changed at the last minute, without consulting him. What action would be the most effective for Jumah?	Tell the manager in charge of shifts that he is not happy about it	Find out if there is some reasonable explanation for the shift changes	2

4	116	Jacob is having a large family gathering to celebrate him moving into his new home. He wants the day to go smoothly and is a little nervous about it. What action would be the most effective for Jacob?	Prepare ahead of time so he has everything he needs available	Accept that things aren't going to be perfect but the family will understand	1
4	117	Song finds out that a friend of hers has borrowed money from others to pay urgent bills, but has in fact used the money for less serious purposes. Song is most likely to feel?	Anger	Contempt	2
4	118	Edna's workmate organizes a goodbye party for Edna, who is going on holidays. Edna is most likely to feel?	Gratitude	Surprise	1
4	119	Kevin has been working at his current job for a few years. Out of the blue, he finds that he will receive a promotion. Kevin is most likely to feel?	Joy	Pride	1
4	120	Rashid needs to meet a quota before his performance review. There is only a small chance that he will be able to do so and there isn't much he can do to improve the outcome. Rashid is most likely to feel?	Hopeful	Scared	2