

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

COGNITIVE AND COMPUTATIONAL TOOLS TO IMPROVE REASONING AND DECISION MAKING IN
THE INTELLIGENCE COMMUNITY

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

DOCTORATE IN COGNITIVE COMPUTER SCIENCE

BY

MATTHEW MARTIN

JANUARY 2026

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

OUTILS COGNITIFS ET INFORMATIQUES POUR AMÉLIORER LE RAISONNEMENT ET LA PRISE DE
DÉCISION DANS LA COMMUNAUTÉ DU RENSEIGNEMENT

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

MATTHEW MARTIN

JANVIER 2026

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de cette thèse se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.12-2023). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

ACKNOWLEDGEMENTS

There are many people who made this work possible and whose help I wish to recognize. First, I would like to thank my research supervisors, Serge Robert and Roger Villemaire, for all their support and guidance for the duration of my PhD.

I would like to thank those who contributed to Chapter 2, particularly my research assistants, Bianca Abbandonato and Aimen Waqar, for their hard work and invaluable contribution to this project. I also wish to thank Tim van Gelder, Ariel Kruger, and other members of the SWARM project team for the data, advice, and technical support that made this project possible.

I wish to thank the co-author of Chapter 3, David Mandel, for his contribution, expertise, advice, and help throughout the project and beyond. I also wish to thank Daniel Irwin for his assistance with this research.

Finally, I wish to thank my parents, Sue and Steve Martin, and my partner, Tess Walsh, for all their love, support and help with all the aspects of the PhD process.

PREFACE

I began the Doctorate of Cognitive Computer Science in September 2019 with the intention of improving institutional decision-making using computational methods. From my current perspective in mid-2026, 2019 seems like a different world, in terms of both institutional decision-making and computational tools. The COVID-19 pandemic and other geopolitical developments highlight the need for decision-making that is rational, transparent, and amenable to democratic oversight and correction.

The world of artificial intelligence has also been transformed by the release of ChatGPT in 2022, making the dream of “strong” or “general” artificial intelligence seem within reach much more quickly than expected. Institutional decision-making will no doubt increasingly incorporate large language models as this technology advances and becomes more widespread. As with human reasoning, the information processing behind the chains of thought these models produce is largely obscure, but the reasoning itself can be subject to the kinds of correctives explored in this thesis. Incorporating large language models with other cognitive and computational tools could increase the rigor and transparency of both human and machine-generated thought. This is an exciting new area for research and practice that I hope to explore in future work.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iii
PREFACE.....	iv
LIST OF FIGURES.....	ix
LIST OF TABLES.....	x
LIST OF ABBREVIATIONS AND ACRONYMS	xii
ABSTRACT	xiii
RÉSUMÉ	xiv
INTRODUCTION.....	1
CHAPTER I	
Theoretical Background and Review of the Literature	7
1.1 Background and structure of the literature review.....	7
1.2 Critique of Structured Analytic Techniques (SATs)	8
1.3 Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE)	13
1.4 Smartly Assembled Wiki-style ARGument Marshalling (SWARM)	15
1.5 Aggregative Contingent Estimation (ACE) program	16
1.6 Algorithmic Recalibration	21
1.7 Discussion	23
CHAPTER II	
The Relationship Between Structured Analytic Techniques And Intelligence Analysis Quality: Evidence From Content Analysis.....	24
2.1 Overview of Chapter 2.....	24
2.2 The Relationship Between Structured Analytic Techniques and Intelligence Analysis Quality: Evidence from Content Analysis	25
2.3 Abstract	26
2.4 Introduction.....	26
2.4.1 CREATE	27
2.5 Methods	29
2.5.1 Participants.....	30
2.5.2 Materials	31

2.5.3	Procedure	35
2.6	Results	37
2.6.1	Reliability Analysis	37
2.6.2	Frequency of Code Usage.....	39
2.6.3	Quality of Reasoning and Accuracy	39
2.6.4	Accuracy	44
2.6.5	Diversity of Codes.....	46
2.7	Discussion	47
2.7.1	Limitations and Future Directions	50
2.7.2	Conclusions.....	52
2.8	Funding details	52
2.9	Disclosure statement.....	52
2.10	Data availability statement.....	52
CHAPTER III		
	Calibration Feedback with the Practical Scoring Rule Does Not Improve Calibration of Confidence.....	53
3.1	Overview of Chapter 3.....	53
3.2	Calibration Feedback With the Practical Scoring Rule Does Not Improve Calibration of Confidence	54
3.3	Citation	55
3.4	Copyright	55
3.5	Abstract	55
3.6	Introduction.....	56
3.6.1	Practical Scoring Rule	58
3.7	Experiment 1	63
3.7.1	Method.....	63
3.7.2	Results	66
3.7.3	Discussion	71
3.8	Experiment 2	72
3.8.1	Method.....	72
3.8.2	Results	73
3.8.3	Discussion	75
3.9	General Discussion	76
3.9.1	Conclusion	79

3.10 Acknowledgments	80
3.11 Conflicts of Interest	80
3.12 Data Availability Statement	81
3.13 Supporting Information	81
CHAPTER IV	
Improving Geopolitical Forecasts with Bayesian Networks.....	82
4.1 Overview of Chapter 4.....	82
4.2 Improving Geopolitical Forecasts with Bayesian Networks.....	83
4.3 Abstract	83
4.4 Introduction.....	84
4.4.1 Bayesian Networks to Predict Geopolitical Outcomes.....	87
4.5 Methods	89
4.5.1 Discretizing the Variables	92
4.5.2 Training the Bayesian Networks.....	94
4.5.3 Logistic Regression Models	94
4.5.4 Model Evaluation	95
4.6 Results	95
4.6.1 Relationships Between Predictor Variables	97
4.6.2 Testing the Assumptions of Logistic Regression.....	99
4.7 Discussion	101
4.8 Conclusions.....	103
4.9 Supplementary Materials	103
4.10 Competing Interests Statement	103
4.11 Financial Support Statement	103
4.12 Acknowledgements	103
4.13 Declaration of Generative AI and AI-assisted Technologies in the Manuscript Preparation Process	104
CHAPTER V	
General Discussion.....	105
5.1 Summary of Results	105
5.2 Interpretation and Implications.....	106
5.3 Limitations	109
5.4 Future Research Directions	110

CONCLUSION	112
APPENDIX A	
Example Stimuli from Chapter 3	114
APPENDIX B	
Dichotomized Logistic Regression Coefficients from Chapter 4	115
SUPPLEMENTARY MATERIALS A	
Supplementary Materials for Chapter 2	116
SUPPLEMENTARY MATERIALS B	
Supplementary Materials for Chapter 3	127
REFERENCES.....	162

LIST OF FIGURES

Figure 1.1. Graphical depictions of noise and bias and how they can be amplified or reduced	11
Figure 1.2. A calibration curve showing overconfidence.....	19
Figure 3.1. Min-max scaled exponentially transformed practical score by block and condition in Experiment 1.	69
Figure 3.2. Min-max scaled exponentially transformed practical score by block and condition in Experiment 2.	74
Figure 4.1. Total mutual information as a function of the number of discrete levels.....	93
Figure 4.2. The Bayesian network structure from Tier 4	98
Figure 4.3. Scatterplot of Absolute Difference from Aggregate and Log Odds of Outcome.	100
Figure A1. Example trial-level feedback for the performance feedback condition in Experiments 1 and 2	114
Figure A2. Example block-level feedback for performance feedback condition in Experiments 1 and 2	114
Figure SB1. Experiment 2 Confidence by Block	137

LIST OF TABLES

Table 2.1. Problem Descriptions for the 2018 and 2020 Challenges.....	33
Table 2.2. Reliability Analysis of Codes.....	38
Table 2.3. Bayesian Regression Models Predicting Quality of Reasoning	42
Table 2.4. Posterior Summaries of Coefficients Most Predictive of Quality of Reasoning	43
Table 2.5. Zero-Order Correlations with Quality of Reasoning	44
Table 2.6. Bayesian Regression Models Predicting Accuracy	45
Table 2.7. Zero-Order Correlations with Accuracy	46
Table 3.1. Experiment 1 Bayesian ANOVA results	71
Table 3.2. Experiment 2 Bayesian ANOVA Results	75
Table 4.1. Performance Metrics for Baseline, Bayesian Network, and Logistic Regression Models	96
Table 4.2. Coefficients from the Tier 4 Lasso Logistic Regression Model with Continuous Variables.....	99
Table B1. Coefficients from the Tier 4 Lasso Logistic Regression Model with Dichotomized Variables .	115
Table SA1. Descriptions of Content Analysis Codes	118
Table SA2. Reliability Analysis of All Codes	121
Table SA3. Posterior Summaries of All Coefficients Predicting Quality of Reasoning.....	123
Table SA4. Zero-Order Correlations of All Predictors of Quality of Reasoning	124
Table SA5. Posterior Summaries of All Coefficients Predicting Accuracy.....	125
Table SA6. Zero-Order Correlations of All Predictors of Accuracy	126
Table SB1. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Block and Condition ANCOVA Results	127
Table SB2. Experiment 2 Min-Max Scaled Exponentially Transformed Practical Score by Block and Condition ANCOVA Results	128
Table SB3. Experiment 1 Bias by Condition and Block ANCOVA Results	129
Table SB4. Experiment 2 Bias by Condition and Block ANCOVA Results	130

Table SB5. Experiment 1 Accuracy by Condition and Block ANCOVA Results	131
Table SB6. Experiment 2 Accuracy by Condition and Block ANCOVA Results	132
Table SB7. Experiment 1 Confidence by Condition and Block ANCOVA Results	133
Table SB8. Experiment 2 Confidence by Condition and Block ANCOVA Results	134
Table SB9. Experiment 2 Confidence Post Hoc Comparisons for Training Condition.....	135
Table SB10. Experiment 2 Confidence Post Hoc Comparisons for Block.....	135
Table SB11. Experiment 2 Confidence Descriptive Statistics	136
Table SB12. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Condition ANCOVA Results.....	138
Table SB13. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Condition ANCOVA Results.....	138
Table SB14. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Condition, Block, and Sex ANCOVA Results	139
Table SB15. Experiment 2 Min-Max Scaled Exponentially Transformed Practical Score by Condition, Block, and Sex ANCOVA Results	140
Table SB16. Experiment 1 Bayesian ANOVA on Min-Max Scaled Exponentially Transformed Practical Score by Condition and Block	146
Table SB17. Experiment 1 Bayesian ANOVA on Bias by Condition and Block	147
Table SB18. Experiment 2 Bayesian ANOVA on Min-Max Scaled Exponentially Transformed Practical Score by Condition and Block	147
Table SB19. Experiment 2 Bayesian ANOVA on Bias by Condition and Block	148
Table SB20. Names of Cities and Countries Used in Experiments 1 and 2	159

LIST OF ABBREVIATIONS AND ACRONYMS

ACE: Aggregative Contingent Estimation

ACH: Analysis of Competing Hypotheses

BF: Bayes Factor

BN: Bayesian Network

CREATE: Crowdsourcing Evidence, Argumentation, Thinking and Evaluation

GJP: Good Judgment Project

IARPA: Intelligence Advanced Research Project Activity

IC: Intelligence Community

QoR: Quality of Reasoning

SAT: Structured Analytic Technique

SWARM: Smartly Assembled Wiki-style ARGument MarShalling

WMD: Weapon of Mass Destruction

ABSTRACT

Improving reasoning and decision-making in the intelligence community (IC) could reduce the chance of wars and other catastrophes. This thesis describes three projects that develop and test cognitive and computational tools to improve reasoning and decision-making. The first project conducted a content analysis of intelligence products to examine the relationships between techniques used, adherence to IC norms, and objective accuracy. Norm adherence and accuracy correlated moderately, and the techniques of Standardizing Probability Ranges and Timeline were predictive of IC norm adherence. In the second project, two experiments used the Practical scoring rule as performance feedback to train better calibration of confidence. The results showed no improvement to calibration from training, contrary to previous results. However, actively open-minded thinking was positively related to overall calibration. The third project used recalibration and weighted aggregation in combination with Bayesian networks (BNs) and logistic regression to improve the accuracy of probabilistic forecasts made during a multi-year forecasting tournament. Recalibration and aggregation alone produced the most accurate forecasts, followed by BNs, then logistic regression. Future studies should use content analysis on a larger set of real intelligence products to address questions of external validity and increase the power of the first project. For the second project, the effects of participant motivation and comprehension should be probed to determine when calibration training with the Practical score is effective. Finally, for the third project, more advanced machine learning techniques like hierarchical BNs and ensemble methods should be explored to further increase the accuracy of recalibrated aggregated forecasts.

Keywords: Reasoning, decision-making, intelligence community, probability, forecasting, content analysis, calibration, machine learning, Bayesian networks, logistic regression

RÉSUMÉ

Améliorer la qualité du raisonnement et de la prise de décision au sein de la communauté du renseignement (CR) pourrait réduire le risque de guerres et d'autres catastrophes. Cette thèse décrit trois projets qui développent et testent des outils cognitifs et informatiques visant à améliorer le raisonnement et la prise de décision. Le premier projet a mené une analyse du contenu des produits du renseignement afin d'examiner les relations entre les techniques utilisées, le respect des normes de la CR et la précision objective. Ce projet a montré que le respect des normes et la précision étaient modérément corrélés, et que l'usage des techniques de standardisation des plages de probabilité et de la chronologie permettaient de prédire le respect des normes de la CR. Dans le deuxième projet, deux expériences ont utilisé la règle de notation pratique comme retour d'information sur les performances afin d'améliorer le calibrage de la confiance. Les résultats n'ont montré aucune amélioration du calibrage suite à la formation, contrairement aux résultats précédemment publiés dans la littérature scientifique. Cependant, la « pensée activement ouverte » était positivement corrélée au calibrage global. Le troisième projet a utilisé le recalibrage et l'agrégation pondérée en combinaison avec des réseaux bayésiens (BN) et une régression logistique afin d'améliorer la précision des prévisions probabilistes réalisées lors d'un tournoi de prévisions pluriannuelles. Le recalibrage et l'agrégation ont produit les prévisions les plus précises, suivis par les BN, puis par la régression logistique. Les études futures devraient utiliser l'analyse de contenu sur un ensemble plus large de produits de renseignement réels afin d'aborder les questions de validité externe et d'augmenter la puissance du premier projet. Pour le deuxième projet, les effets de la motivation et de la compréhension des participants devraient être étudiés afin de déterminer quand la formation au calibrage avec la règle de notation pratique est efficace. Enfin, pour le dernier projet, des techniques d'apprentissage automatique plus avancées, telles que les BN hiérarchiques et les méthodes d'ensemble, devraient être explorées afin d'accroître encore la précision des prévisions agrégées recalibrées.

Mots clés: Raisonnement, prise de décision, communauté du renseignement, probabilité, prévision, analyse de contenu, calibrage, apprentissage automatique, réseaux bayésiens, régression logistique

INTRODUCTION

On September 11, 2001, two airplanes crashed into the World Trade Center towers in New York City, killing nearly 3000 people. There had been warnings from various members of the United States (US) Intelligence Community (IC) that Osama Bin Laden was plotting a terrorist attack on the US, but little serious action was taken in response. Intelligence agencies were apparently caught off-guard by the method of the attack, since airplanes had not previously been used as a weapon of mass destruction (National Commission on Terrorist Attacks Upon the United States, 2004). In response to the terrorist attack, the US invaded Afghanistan and later Iraq in its “War on Terror”. Iraq was invaded under the premise that it possessed weapons of mass destruction (WMD) and posed danger to the US. The war in Iraq lasted over ten years and cost the lives of over 5000 Americans, 100,000 Iraqis, and \$4 trillion USD (Chang et al., 2016; Crawford, 2013). It later became clear that Iraq did not possess WMDs, thus invalidating a primary justification for the war. A later bipartisan investigation into the causes of the Iraq war suggested that the IC had made overconfident assessments of the presence of WMDs and some policymakers had politicized the available evidence (Chang et al., 2016; Select Committee on Intelligence, 2006).

The investigations undertaken in the wake of 9/11 and the Iraq war prompted a series of reforms of the US IC, which included funding research to improve its tradecraft (Chang et al., 2016; National Commission on Terrorist Attacks Upon the United States, 2004). This thesis investigates how cognitive and computational tools can improve the reasoning and decision-making of institutions such as the IC. Two of its chapters build directly on work funded by the US IC. While the applications of this research are primarily aimed at the IC, its methodologies and insights are more generally relevant to reasoning under uncertainty and can be applied to many other domains.

The IC was chosen as the primary domain of application for this thesis for three reasons. First, the IC faces many difficult epistemic problems, as it often requires making decisions under profound uncertainty. Second, its decisions have dramatic effects on the world, such as causing or preventing wars, so improving its decision-making capabilities can be highly impactful. Third, the domain has been relatively neglectful in implementing evidence-based and scientifically sound methodologies in its practices. As is elaborated in the literature review in Chapter 1, many of the techniques used by the IC were created on an ad hoc basis and have not been subject to sufficient empirical scrutiny, so this thesis seeks to improve the status quo by evaluating and developing improved reasoning and decision-making methodologies (Chang et al., 2018).

Governments, militaries, and businesses make innumerable decisions in a chaotic and changing world. Unfortunately, unaided human reasoning is often flawed and biased, as documented by Kahneman (2011) and others. In the context of this thesis, bias refers to systematic error or violations of norms of reasoning. These include both epistemic and institutional norms. Epistemic norms include the norms of logic, probability theory, and mathematics. Epistemic norms also include other practices related to forming accurate beliefs, such as considering a wide range of information and being willing to change one's beliefs when given good reason to (discussed further below). Institutional norms relate to good reasoning practices held by organizations, such as communication standards ensuring transparency and oversight.

Probability theory plays an important role in this thesis, as it is the main normative and theoretical framework by which the quality of reasoning under uncertainty is judged. It also forms part of the methodology of each of the three main body chapters. Probability theory can be viewed as an extension of classical logic. Where classical logic is bivalent (representing truth values with only 'true' or 'false') and assertoric (expressing only statements of fact), probability theory is multivalent (expressing degrees of uncertainty in real values between 0 and 1) and can express more than mere statements of fact, such as subjective confidence in future possibilities (Demey et al., 2019). Probability theory has a precise syntax for performing logical operations that preserve the probabilities of the premises, making it highly expressive and a valuable tool for representing reasoning under uncertainty. The fields of statistics, artificial intelligence, and cognitive science have used probability theory extensively in representing rational deliberation with uncertainty (Demey et al., 2019).

Three important concepts in probability theory are unconditional probability, conditional probability, and Bayes' theorem. Unconditional probability expresses the probability of a variable, e.g., A , being in a particular state, without consideration of the states of other variables, and is expressed as $P(A)$. Conditional probability expresses the probability of a state of a variable, e.g., A , given the state of another variable, e.g., B , and is expressed as $P(A/B)$. The probabilities of A and B can inform each other via Bayes' Theorem, described in (1),

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad (1)$$

where $P(A/B)$ and $P(B/A)$ represent conditional probabilities, referred to as the posterior probability and the likelihood, respectively. $P(A)$ and $P(B)$ are unconditional probabilities, referred to as the prior and

marginal probability, respectively. Bayes' theorem allows for a mathematical description of both inference and learning. If A is a hypothesis, inference is performed by updating the prior probability of the hypothesis, $P(A)$, to the posterior probability, $P(A/B)$, after observing some evidence, B , as expressed as the likelihood and marginal probabilities of B , $P(B/A)$, and $P(B)$, respectively. The field of Bayesian statistics uses Bayes' theorem in this way to estimate the probability of hypotheses given a set of evidence (Kruschke, 2015).

Learning can be modeled using Bayes' theorem by iteratively substituting the prior probability at one time, t_n , with the posterior probability from a previous time, t_{n-1} . If there has been any change to the state of the evidence between the two observations, then the posterior probability at time t_n will be updated to reflect a combination of the new and old information (Shanks, 2006). Bayes' theorem plays a role as a normative account of reasoning in Chapter 2 of the thesis, where the correct answer to some problems posed to intelligence analysts was determined using Bayes' theorem.

In addition to epistemic norms like those of probability theory, improving institutional decision-making also implies improving adherence to institutional norms important to the organization in question. For example, the US IC lists nine analytic standards that serve as norms about quality of reasoning and effective communication that all analytic products are expected to follow (Office of the Director of National Intelligence, 2015a).¹ The standards for analytic products are:

1. Properly describes quality and credibility of underlying sources, data, and methodologies
2. Properly expresses and explains uncertainties associated with major analytic judgments
3. Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments
4. Incorporates analysis of alternatives
5. Demonstrates customer relevance and addresses implications
6. Uses clear and logical argumentation
7. Explains change to or consistency of analytic judgments (with those presented in previous reports)
8. Makes accurate judgments and assessments

¹ I use the term *norm* here to refer to prescriptive or injunctive norms, i.e., ideals held by the institution. It does not refer to descriptive or social norms, nor does it imply that these ideals are typically implemented in practice.

9. Incorporates effective visual information where appropriate (van Gelder et al., 2020, p. 206)

Note that in addition to the epistemic norms of “Uses clear and logical argumentation” and “Makes accurate judgements and assessments”, many of the requirements are primarily related to clear communication and avoiding misinterpretation. This list represents a different normative account of reasoning from that of say, probability theory, in that this set of norms is qualitative and primarily about using language that does not misinform the audience. Probability theory would not be able to represent most of these norms. Adherence to the above norms and their relationship to objective accuracy, as assessed partly through probability theory, are the focus of Chapter 2 of this thesis, as discussed in more detail below.

Violations of the norms of probability theory are not the only form of epistemic bias relevant to this thesis. Other epistemic biases, such as status quo bias, can negatively affect rational deliberation. This bias describes the tendency to over-weight the hypothesis that things will stay the same and neglecting to consider new, unexpected changes (Chang et al., 2018). It also describes the preference for choosing options that do not require making changes, even if a new option might be better (Samuelson & Zeckhauser, 1988). Status quo bias was arguably at work in the failure to prevent the 9/11 terrorist attacks, as existing protocols and intelligence assumptions were ill-equipped to handle the new possibility of passenger airplanes being used as a weapon (National Commission on Terrorist Attacks Upon the United States, 2004; Shapira, 2025). Probability theory does not provide guidance on which hypotheses or pieces of data are appropriate to consider when making a decision. It can only provide a syntax for assigning and updating probabilities associated with the hypotheses under consideration, given the data being examined. Meta-questions related to which hypotheses and data to consider are epistemological questions over and above those of probability theory. These questions are addressed with different epistemological norms that nonetheless play an important role in decision-making.

Aside from status quo bias, two other cognitive biases important to this thesis are confirmation bias and overconfidence. Confirmation bias is the tendency to attend to or otherwise favour evidence that supports a conclusion that one already believes or wishes to be true, rather than seek out and appropriately weigh disconfirming evidence. Like with status quo bias, confirmation bias can manifest itself as a failure to consider enough alternative hypotheses (Nickerson, 1998). Overconfidence is a bias that describes the tendency for people to systematically overestimate their own knowledge or ability. For example, someone

consistently claiming to be 100% confident in a proposition or prediction, when in fact they are only correct 80% of the time when making such claims, would be demonstrating overconfidence. All three of these biases have been the targets of intervention on the part of the IC to improve its tradecraft (Chang et al., 2018).

The IC has primarily tried to address these kinds of biases using structured analytic techniques (SATs), which are formalized methodologies intended to combat bias and make analysis more rigorous and transparent (Chang et al., 2018). For example, the US Government's document *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis* (henceforth *Tradecraft Primer*) describes 12 SATs that are designed to mitigate bias, 10 of which target status quo bias (Chang et al., 2018; US Government, 2009). One SAT called analysis of competing hypotheses targets confirmation bias, as it forces analysts to explicitly indicate the degree to which each relevant piece of evidence is consistent or inconsistent with each hypothesis under consideration (Chang et al., 2018). Overconfidence is also targeted by SATs in the *Tradecraft Primer*, although no specific SAT is explicitly linked to it (US Government, 2009). The extent to which techniques such as SATs improve quality of reasoning and accuracy is reviewed in Chapter 1 and investigated empirically in Chapter 2 of this thesis. Chapter 2 is a version of an article submitted to the journal *Intelligence and National Security*.

Another proposed method of mitigating of overconfidence (and the inverse bias of underconfidence) is through calibration training, where analysts are trained to align their subjective confidence with their objective accuracy (Chang et al., 2018). Rather than relying on analytic techniques which may exacerbate bias in one direction or another (see [Chang et al. \(2018\)](#) and Chapter 1 for a discussion of how this may happen), calibration training using a scoring rule that incentivizes reducing both overconfidence and underconfidence may be an effective way to improve the accuracy of subjective probability judgements. While SAT methodologies attempt to prevent analysts from using intuitive judgements that are assumed to be biased, calibration training would seek to improve the intuitions of the analysts so that they are less biased to begin with (Chang et al., 2018). There have been several experimental demonstrations of successful calibration training, although most of these have involved very long training sessions or specialized participants and often use unintuitive scoring methods that may inhibit learning (see Chapter 1 for a review). Chapter 3 of this thesis investigates whether calibration can be improved in a short-duration training session using an intuitive scoring rule meant to facilitate learning. Chapter 3 is a version of an article published in the journal *Futures & Foresight Science* (Martin & Mandel, 2025).

A separate methodology for reducing the impact of biases like under- and overconfidence is to correct for bias post hoc using computational means. This methodology has been used primarily in the context of probabilistic forecasting, where multiple forecasts about future events can be aggregated and recalibrated to create a more accurate combined forecast (Ungar et al., 2012). Chapter 4 of this thesis explores Bayesian Networks (BNs) as a means of increasing the accuracy of forecasts, along with logistic regression and a form of recalibration and aggregation. Bayesian networks are a technique from the field of artificial intelligence based in probability theory that represent relationships between variables as conditional probabilities (Canonaco et al., 2025). Given their ability to represent complex conditional relationships between variables, and that discrete BNs make fewer parametric assumptions about the data than other methodologies like logistic regression, BNs may prove better than other existing methodologies for creating algorithmically enhanced forecasts. Chapter 4 is a version of an article submitted to the *International Journal of Forecasting* that investigates this question.

In sum, the thesis proposes and evaluates cognitive and computational aids to assist rational decision-making under uncertainty via adherence to probability theory and other epistemic and institutional norms. Chapter 1 reviews the literature relevant to the three studies of the current thesis, with a focus on the previous projects conducted by the IC to improve its tradecraft that this thesis builds on. Chapter 2 evaluates a set of intelligence products and examines the extent to which using SATs and other techniques predicts objective accuracy and adherence to IC norms. Chapter 3 assesses the effectiveness of a training paradigm to improve the calibration of subjective confidence and accuracy of participants' beliefs based on a novel scoring rule that is easily interpretable and incentivises accurate calibration. Chapter 4 explores the effectiveness of BNs as a means of making accurate geopolitical forecasts, in conjunction with recalibration and aggregation methods applied to crowd-sourced forecasting data. The General Discussion and Conclusion synthesize the findings and suggest areas for future research.

CHAPTER I

Theoretical Background and Review of the Literature

1.1 Background and structure of the literature review

This thesis builds on two research projects funded by the US intelligence community (IC): the Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE) program and the Aggregative Contingent Estimation (ACE) program. The CREATE program is related to Chapter 2 of the thesis, which analyses a dataset of intelligence products produced during this program and seeks to answer the question of which analytic techniques found in the products are most predictive of high-quality reasoning and accuracy. The ACE program is relevant to Chapters 3 and 4 of this thesis. The ACE program sought to determine the best means of eliciting and aggregating probabilistic forecasts about future geopolitical events from a large group of forecasters (IARPA, 2010). Chapter 3 tests a training paradigm to improve the calibration of subjective confidence and accuracy, a component of forecasting skill. Chapter 4 analyzes a dataset collected during the ACE program using Bayesian networks (BNs) and recalibration and aggregation to improve aggregate forecast accuracy.

The literature review in this chapter first outlines the context behind the CREATE program and provides a critical review of the use of structured analytic techniques (SATs) within the IC. The CREATE program is then introduced in more detail as a continuation of the IC's interest in developing and testing SATs as a means of reducing bias in analysis. The Smartly-assembled Wiki-style Argument Marshalling (SWARM) project from the CREATE program is explained in more detail, as it is the basis of the research in Chapter 2. The theory, methodology, and results of the SWARM project are described with the aim of setting the context and justifying the follow-up research conducted in Chapter 2. To briefly recapitulate the arguments in the literature review, numerous authors have critiqued the status quo of SAT use in IC practice on the grounds that the techniques have not undergone the level of empirical scientific scrutiny that would warrant their widespread adoption in the IC, and there are reasons to believe that some SATs have detrimental effects on reasoning and decision quality. The study in Chapter 2 analyses the content of the intelligence reports written during experiments simulating IC conditions and codes for the kinds of techniques used in them to determine which techniques are related to measures of high-quality reasoning and objective accuracy. This kind of empirical analysis of the relationship between techniques used in

realistic analysis settings and metrics of reasoning quality and accuracy is relatively rare and is particularly important given the critiques of SAT use elaborated below.

Following the literature review related to the CREATE program, the ACE program is reviewed, providing important context related to forecasting, calibration, and the psychological variables behind it. This leads to a discussion of calibration training, which is the focus of Chapter 3. The literature review then discusses algorithmic means of aggregating and recalibrating forecasts to increase accuracy, which was a key element of the ACE project and the focus of Chapter 4. In Chapter 4, BNs are used in conjunction with recalibration and aggregation to determine if they can increase aggregate accuracy compared to other methods including logistic regression. In the next section, the status quo of SATs and their use in the IC is discussed and critiqued, which provides the context behind Chapter 2.

1.2 Critique of Structured Analytic Techniques (SATs)

Structured analytic techniques are tools designed to make analysts' thought processes more transparent and logically sound by reducing bias and organizing complex evidence, thus increasing accuracy compared to unaided analysis. In particular, they are meant to ameliorate status quo bias and confirmation bias, among others (Chang et al., 2018). They have been used within the IC for decades, are an important part of analyst training, and are even considered necessary for high-quality analysis (although in practice they are used somewhat rarely, as discussed below (Artner et al., 2016)). The only official list of SATs published by the US IC itself is the *Tradecraft Primer* (US Government, 2009). Another influential work is *Structured Analytic Techniques for Intelligence Analysis* (Pherson & Heuer, 2021), which describes how to use 66 different techniques.

However, there are four main criticisms in the research literature about how SATs are used in the IC. The first is that most of the SATs endorsed by the IC have not been rigorously evaluated for their effectiveness in mitigating bias or increasing accuracy or transparency (Artner et al., 2016; Coulthart, 2017). The second criticism is that some of the techniques championed by the IC, such as analysis of competing hypotheses (ACH), have been demonstrated to be ineffective or detrimental in several studies (Dhami et al., 2019; Karvetski et al., 2020; Karvetski & Mandel, 2020; Mandel et al., 2018; Wilcox & Mandel, 2024). The third criticism is that many of the SATs endorsed by the US IC show methodological and theoretical flaws that would lead to the expectation that they actually increase bias (Chang et al., 2018; Mandel & Tetlock, 2018). The fourth criticism is that the predominant method of analysis remains intuitive, subjective judgements

despite some efforts to increase the IC's use of evidence-based techniques in its tradecraft (Artner et al., 2016; Coulthart, 2016; Mandel & Tetlock, 2018). Each of these points are detailed in turn, below.

Despite their promotion by institutions in the US IC, empirically testing how effective SATs are at reducing bias and improving analysis has been relatively neglected. According to interviews with a selection of US IC managers and practitioners in 2014, neither how often SATs were being used, nor their effect on analysis quality were routinely measured by any US intelligence agency (Artner et al., 2016). Subsequently, a systematic review by Coulthart (2017) found that only six of the 12 techniques described by the *Tradecraft Primer* had ever been evaluated empirically in the published literature. Furthermore, only three were considered to have evidence of 'credible' quality: devil's advocacy, ACH, and brainstorming. The most promising technique was devil's advocacy, with three quarters of studies finding that it improved accuracy or effectively critiqued assumptions. Analysis of competing hypotheses reduced confirmation bias in some circumstances, but these findings were inconsistent for different types of groups and different measures of bias. Similarly, brainstorming also showed mixed results, as it can actually reduce the quality and quantity of ideas when used in a face-to-face context (Coulthart, 2017). These findings suggest that the status quo in the IC often ignores the power of the scientific method and may be liable to waste resources on adoption of useless or even counterproductive techniques. Since the most reliable way to improve analytic tradecraft is to rigorously measure and compare the effects of interventions trying to improve it, failing to do so undermines the stated goals of the SATs themselves: to produce transparent, unbiased, and accurate analysis.

The ACH technique provides a useful example of how poorly designed analytic techniques can potentially harm analysis. The ACH technique is meant to help analysts compare the plausibility of multiple hypotheses under consideration. By design, it requires analysts to find and list all reasonable alternative hypotheses that could explain a given body of evidence, instead of focusing on a single preferred hypothesis, thus avoiding confirmation bias. The ACH technique involves constructing a two-dimensional matrix where all hypotheses that could explain the data are arranged in columns and the relevant pieces of evidence are listed in rows. Analysts then fill in the matrix with judgements of whether each piece of evidence is consistent, inconsistent, or not applicable for each hypothesis. Finally, each hypothesis is assessed with a total inconsistency score; the sum of the pieces of evidence that are deemed inconsistent with the hypothesis. The hypothesis with the lowest inconsistency score is then supposed to be the most plausible (U.S. Government, 2009).

The ACH's methodology has received both theoretical and empirical critiques. On theoretical grounds, it has been criticized for the ambiguity of its procedure and its violation of the norms of probability theory. The technique fails to clearly define how analysts should break down interconnected pieces of evidence and hypotheses into their component parts. It also fails to provide clear instructions on how to evaluate whether a piece of evidence is consistent or inconsistent with each hypothesis. Expressed as probabilities, it is unclear whether an inconsistent piece of evidence is one where $P(\text{Evidence}|\text{Hypothesis})$ is low, or $P(\text{Hypothesis}|\text{Evidence})$ is low, since the technique does not disambiguate between the two.

Regarding its violation of probability theory, ACH does not integrate any evidence that is judged to be consistent with a hypothesis. The technique apparently assumes a Popperian interpretation of scientific falsification, where scientific hypotheses can only ever be disconfirmed and not confirmed (Mandel, 2019). However, a Bayesian understanding of hypothesis testing based in probability theory would imply that calculating the posterior probability of a hypothesis involves integrating its prior probability and the probabilities of all pieces of evidence that are both consistent and inconsistent with the hypothesis, as expressed in the likelihood and marginal probabilities. Ignoring consistent evidence therefore biases the method away from conclusions that would follow from a Bayesian analysis grounded in probability theory (Mandel, 2019). These shortcomings of ACH could lead to inconsistent application of the method and possible introduction of systematic bias by ignoring evidence consistent with hypotheses (Chang et al., 2018). This expectation appears to be borne out empirically, as a recent review of ACH concludes that it offers little benefit and may be detrimental to analysis (Wilcox & Mandel, 2024). This ultimately leads the authors to recommend against its mandatory training and use in the IC unless future studies substantially override the existing findings. Similar criticisms apply to other SATs promoted by the IC, which ignore the bipolar nature of bias and the effect of adding noise (Chang et al., 2018).

In a more wide-ranging critique of the effectiveness of SATs in general, Chang et al. (2018) argue that judgement error stems from two sources: random noise and systematic bias. Figure 1.1 shows graphical representations of noise and bias. Noise is random error (Figure 1.1A), whereas systematic bias is any consistently directional error (Figure 1.1C). Chang et al. (2018) propose that many SATs misunderstand bias to be unipolar, where reducing bias consists of moving judgement in the opposite direction of the bias (Figure 1.1E). For example, if overconfidence were misconstrued this way, the prescribed mitigation for general overconfidence would be to encourage or train people to systematically lower their confidence. However, this bias, like many others, are bipolar in nature, so a miscalibrated mitigation could lead to over-

correction and the inverse error (Figure 1.1F). In some situations, judgements have been shown to be systematically underconfident, such as in the hard-easy effect, where people tend to be underconfident with easy tasks, and over-confident with hard tasks (Juslin et al., 2000). Underconfidence may also be situational or cultural, as it has been observed in Canadian intelligence officers, where it could be due to factors such as job experience, accountability pressures, and risk aversion (Mandel & Barnes, 2018). Effective debiasing strategies need to reduce error by mitigating bias in any direction. For confidence debiasing, calibration feedback, where participants receive feedback about their direction of error (discussed more in Chapter 3), is sensitive to both directions of bias and is therefore a better mitigation than systematically reducing confidence (Chang et al., 2018).

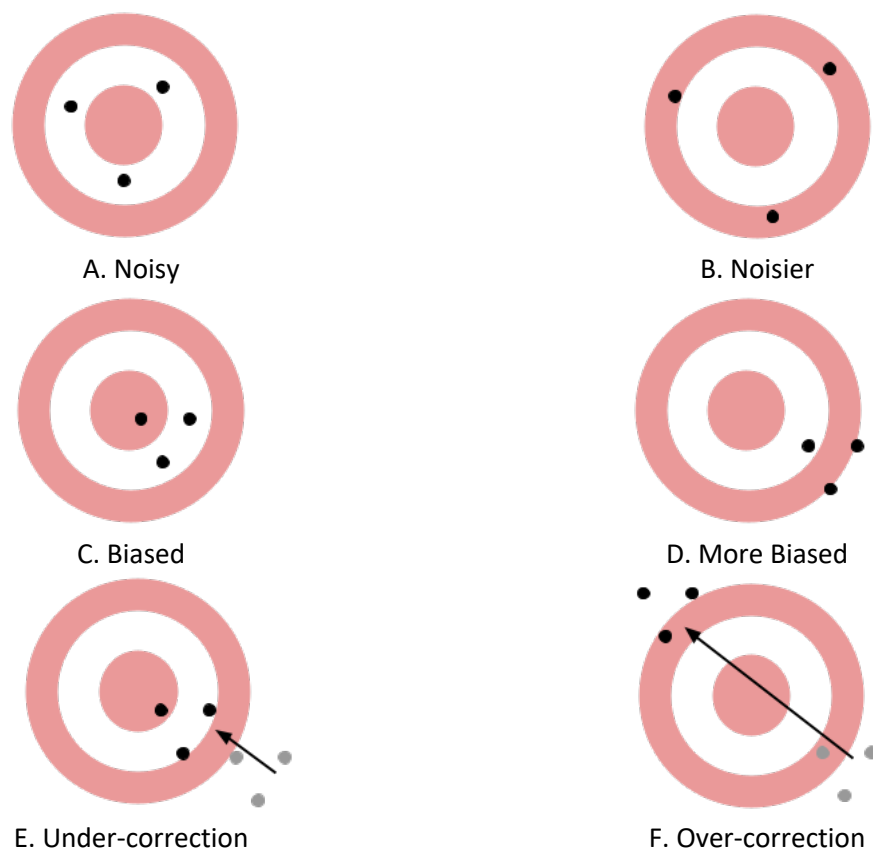


Figure 1.1. Graphical depictions of noise and bias and how they can be amplified or reduced

Chang et al.'s (2018) second criticism is that SATs have a tendency to ignore the effect of noise in judgement error. They contend that it is often assumed that following a step-by-step process makes thinking more transparent and reliable, thereby reducing noise. However, many SATs make analysts decompose problems into sub-components, which introduces many subjective judgements into the analysis. For example, ACH requires analysts to decompose a body of evidence into individual rows and

rate the consistency of each piece of evidence with each hypothesis, while providing unclear information on how to implement these steps. Requiring analysts to complete an unclear procedure with multiple steps still leaves the door open to bias, and may increase noise, since there are more subjective decisions to make when decomposing a problem into its components. This was demonstrated in exercises involving over 3000 students from U.S. intelligence agencies, where Heuer and Pherson (2014) found that “students disagree on how to rate the evidence in about 20-30 percent of the cells in [an ACH] matrix.” (p. 311). This finding highlights that the ACH technique does not define its terms rigorously enough to be applied consistently. Chang et al (2018) recommend that probability theory should be incorporated into SATs so that the analysts’ reasoning can be better verified for accuracy and consistency. They also advocate for substantiating that SATs are effective at reducing both bias and noise by examining their reliability across time and across different groups of analysts.

Regarding the fourth critique that SATs are underutilized in IC practice, some authors have commented that the status quo of intelligence analysis “lacks many of the scientific community’s self-correcting features” (Cooper, 2005, pp. 5-6), and rather consists of a “collection of suggestions and tips for writing and communicating with co-workers, supervisors, and consumers instead of focusing on a thorough examination of the analytic process and techniques” (Johnston, 2005, p 17; see also Mandel and Tetlock (2018)). For example, to determine what proportion of intelligence products used SATs in their analysis. Artner et al. (2016) reviewed a sample of unclassified documents produced by the CIA, National Intelligence Council (NIC), and Defence Intelligence Agency (DIA). They found that 23 / 29 (79%) CIA intelligence assessments, 10 / 14 (71%) NIC analytic products, and 8/20 (40%) documents from a mixed DIA and CIA sample showed no evidence of SAT use. Of the 22 documents that did use SATs, the analyses generally showed more consideration of alternative explanations, distinguished between assumptions vs judgements more effectively, and had better logical argumentation than documents without SATs. The authors note that analysts often view SATs as too cumbersome and time-consuming to be useful by analysts, particularly on short-deadline projects. However, the small sample size and qualitative nature of Artner et al.’s (2016) review prevents drawing strong conclusions. In a similar vein, a different survey of 80 analysts at the US State Department’s Bureau of Intelligence and Research found that 33% of analysts reporting never using SATs in their analyses, 46% report rarely using them, and only 21% of analysts use them ‘sometimes’. Reluctance to use methods of unknown efficacy and lack of training with them were the main barriers to uptake (Coulthart, 2016).

Despite the criticisms of the US IC status quo outlined above, changing to a new state of affairs where analytic techniques are rigorously evaluated using scientific theory and empiricism is both desirable and possible.

1.3 Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE)

Partly to address these critiques of the status quo of SAT use in the IC, IARPA invested in the CREATE program which investigated whether incorporating crowdsourcing and SATs would improve the rigor and quality of reasoning in analytic products. CREATE was a competition between four teams of researchers from four different universities. The Bayesian Argumentation via Delphi (BARD) project, of Monash University, used a combination of the Delphi technique paired with BNs to help teams reach agreement about probabilistic hypotheses (Nicholson et al., 2019). The Co-Arg team, from George Mason University, developed a cognitive assistant that used Wigmorean inference networks and incorporated a theory of evidence-based reasoning (Boicu et al., 2018). The Trackable Reasoning and Analysis for Crowdsourcing and Evaluation (TRACE) project, of Syracuse University, used behavioural nudges to guide analysts towards the most relevant information and established SATs for their analytical context (Stromer-Galley et al., 2021). Finally, the Smartly-assembled Wiki-style Argument Marshalling (SWARM) project, of the University of Melbourne, implemented an online collaboration platform that sought to promote engagement, collaboration and natural expertise among a large cohort of analysts (van Gelder et al., 2018). These four systems were compared to two control groups who only had access to Google docs for their analyses, one control group was for teams and one was for individuals (Boicu et al., 2018).

The four systems differ in the quantity of training needed for effective usage, as well as how constraining the tool is on the analyst. The BARD and Co-Arg platforms both require substantial training of the user before they can effectively operate the system. In the case of BARD, participants must produce BNs using a Delphi technique for their analysis. Training on this system takes approximately four hours of training in Bayesian reasoning and its implementation. For the Co-Arg platform, analysts need hours of training to learn a theory of evidence-based reasoning and how to create Wigmorean networks using the system (Boicu et al., 2018).

In comparison, the TRACE and SWARM platforms have less training requirements, preferring flexibility and positive user experience over unfamiliar formal techniques. The TRACE platform guides its analysts to use a set of well-known SATs in their reporting, including structured debate, key assumptions check, and ACH.

It also provides a report template for structuring analyses according to the norms of high-quality reasoning promoted by the IC. The system also nudges analysts to use other optional tools, aiding the user at key points of the analysis (Stromer-Galley et al., 2021).

Compared to TRACE, the SWARM platform can be said to provide an even greater degree of flexibility to users. When using SWARM, analysts are encouraged, though not required, to incorporate a set of “logical lenses” in their analysis, drawn from the “lens kit” provided by the SWARM team (van Gelder et al., 2020). The lens kit is an online resource that contains articles about a wide variety of intelligence-related topics in a style similar to Wikipedia. The range of topics includes an introduction to probability theory and Bayes theorem, BNs, confirmation bias, game theory, ACH, and tips for effective use of visual information, along with more concepts and techniques. In structure, the SWARM platform resembles popular online question-and-answer forums like Stack Overflow, a site where users can submit any coding-related questions, or answer the questions of other users in exchange. Users also engage in upvoting and downvoting others’ responses to any given question, to sort them by quality, and raise the best answer to the attention of the asker. This structure leverages the wisdom of the crowd and represents the key feature that SWARM seeks to replicate with its interactive platform.

With SWARM, the types of information shared on the platform include *Resources*, which can be individual pieces of information or partial analyses, and *Reports*, which are candidates for final submission. Users discuss, dispute, and collectively rank the submitted resources and reports. The most highly rated report rises to the top and is considered the group consensus. The SWARM platform provides a collaborative and engaging environment where users have the freedom to complete problems as they see fit. The lens kit offers users a set of conceptual frameworks to deploy and allows them to choose which ones may or may not be useful for the task at hand. Users rely on each other to assess their reasoning and their use of the logical lenses (Sinnott, 2019; van Gelder et al., 2018, 2020).

The BARD, TRACE, and SWARM platforms have all undergone empirical testing of their ability to enhance quality of reasoning. Compared to unaided control groups, these three systems demonstrate large, statistically significant improvements (see: Korb et al., 2020; Stromer-Galley et al., 2021; van Gelder et al., 2020). The Co-Arg research team has not published an empirical analysis of the system’s effectiveness. Unfortunately, the published research about the three empirically validated systems from the CREATE program does not allow for direct comparison between the systems, as their experimental conditions were

incomparable. While all three platforms have demonstrated their effectiveness at improving analysis quality, the reports generated using the SWARM platform arguably provide the most valuable dataset for follow-up research, as argued next.

1.4 Smartly Assembled Wiki-style ARGument Marshalling (SWARM)

The SWARM platform underwent empirical testing in both in 2018 and 2020, showing promising results. In the 2018 Challenge, four different questions simulating real-world intelligence problems were presented to 24 teams of analysts. The teams were formed from three distinct populations. The first consisted of participants recruited from the public, the second consisted of analysts (such as risk or policy analysts) who did not work for government bodies, and the third consisted of trained analysts recruited predominantly from government organizations in Australia and New Zealand. For one of the four problems, the groups of trained government and non-government analysts were asked to use their standard tools for analysis, rather than the SWARM platform. Overall findings showed the SWARM system was positively reviewed by its users and the tool enhanced the quality of reasoning in their reports (van Gelder et al., 2020). A similar test was conducted in 2020 (i.e., the 2020 Challenge) with different teams of participants recruited from both the public and from government organizations (Thorburn et al., 2020). This test focused on the relationship between accuracy and quality of reasoning, as elaborated below (Hunt Laboratory & University of Melbourne, 2020).

All of the reports in the 2018 and 2020 Challenges were given a Quality of Reasoning (QoR) score using a version of the Rating Scale for Evaluating Analytic Tradecraft Standards² (Hunt Laboratory & University of Melbourne, 2020; Office of the Director of National Intelligence, 2015b; van Gelder et al., 2020). The IC Rating Scale quantifies an intelligence product's degree of alignment with the analytic tradecraft standards of the US IC (Office of the Director of National Intelligence, 2015a). Two pages of guidance for each of the standards are provided in the IC Rating Scale, including instructions for evaluating a report as Poor, Fair, Good, or Excellent. Standard #7 ("Explains change to or consistency of analytic judgments") was not scored in the 2018 and 2020 Challenges because it wasn't relevant to the experimental setting (van Gelder et al., 2020). On the remaining eight standards, each report received a score of 1 to 4 (from Poor to Excellent)

² This thesis uses the term Quality of Reasoning score for consistency with the original research by van Gelder et al. (2020). It could be argued that this term is a poor descriptor of the aggregate score of the IC Rating Scale, as most of the standards do not measure reasoning as commonly understood in the cognitive science literature. Thorburn et al. (2022) use the term IC Rating Scale Score for this measure, which may be a better descriptor.

with the sum of eight scores yielding the overall QoR score ranging from 8 to 32 points. Three independent raters scored each report, and the average of the three was used as the report's final QoR score.

A Bayesian probability puzzle with an objectively verifiable answer was included in the 2018 Challenge (van Gelder et al., 2020). All four of the problems in the 2020 Challenge had objectively verifiable answers, including another Bayesian probability puzzle (Hunt Laboratory, 2019; Thorburn et al., 2020). Thus, the correspondence between QoR scores and objective accuracy can be evaluated both when probability theory is directly applicable and in other contexts where probability theory is less relevant. It also enables an analysis of how SATs and other methodologies used in the reports relate to both QoR scores and accuracy for five of the eight questions.

The original authors of both challenges did not perform a detailed examination of which elements of SWARM were related to higher QoR scores and objective accuracy. The SWARM platform allowed participants to freely choose which techniques to use in their analyses, better reflecting naturalistic conditions compared to the other CREATE systems, which put more constraints on users. SWARM's lens kit provides a useful framework to categorize the techniques used by the analysts. This would enable a fine-grained examination of the relationships between QoR scores, accuracy, and techniques used in the reports. Chapter 2 of this thesis uses a content analysis methodology to analyze the SWARM reports and pursue this line of research.

1.5 Aggregative Contingent Estimation (ACE) program

The second IARPA-funded research project that this thesis draws from is the ACE program. The ACE program was a multiyear competition running from 2010 to 2015 which sought to identify the most effective methods of eliciting and aggregating participants' probabilistic forecasts about future geopolitical events. The investigation also examined the impacts of individual differences like personality factors and measures of IQ, situational variables, such as forecasts coming from varying team structures, and interventions impacting forecasting accuracy, like participant training programs. Chapter 3 draws on the idea of a training intervention with the potential to usefully boost forecasting skills. It presents a short calibration training paradigm that is intended to align subjective confidence ratings with long-run accuracy, an important component of accurate forecasting. If successful, the intervention might form the basis of a useful forecasting training program. A second core element of the ACE program was applying aggregation algorithms to increase predictive accuracy. This line of research is pursued in Chapter 4 using a dataset

from the ACE program and employing BNs alongside recalibration and aggregation to generate more accurate forecasts. To set the context for new research, the ACE program is described next, followed by an overview of calibration, then algorithmic recalibration and aggregation.

The Good Judgement Project (GJP) headed by Philip Tetlock and Barbara Mellers of the University of Pennsylvania was the winner of the ACE program (Good Judgment Inc., 2025). They asked over a thousand forecasters to give probabilistic predictions about the outcomes of hundreds of geopolitical events. Each prediction was provided as a subjective, two-decimal probability of the event occurring (e.g., there is an 80% chance that Bashar al-Assad will remain President of Syria through January, 31, 2012). Participants' Brier scores were used to measure accuracy (Brier, 1950; Ungar et al., 2012).

The Brier score is one example of a proper scoring rule. A proper scoring rule has the property that a forecaster must report the true probability distribution to maximize their score. Forecasters are thus incentivized to be as honest and well-calibrated as possible when their subjective probability estimates are scored with a proper rule (Gneiting & Raftery, 2007). The Brier score (BS) is given in (1), below:

$$BS = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2 \quad (1)$$

where N is the total number of forecasts; f_i is the probability assigned to the i th forecast; and o_i is the outcome of the i th event, (0 or 1). The Brier score is conceptually similar to mean squared error, where scores greater than zero indicate more deviation from perfect predictive accuracy. The Brier score is affected by both forecaster performance and the unpredictability of the forecasting environment. To distinguish the sources of error, the Brier score is often decomposed into three components: a variance index (VI), discrimination index (DI), and calibration index (CI; Mandel & Barnes, 2014). The Brier score decomposition is given in (2), with the variance index in (3), the discrimination index in (4), and calibration index in (5), below:

$$BS = VI - DI + CI \quad (2)$$

$$VI = \bar{o}(1 - \bar{o}) \quad (3)$$

$$DI = \frac{1}{N} \sum_{k=1}^K N_k (\bar{o}_k - \bar{o})^2 \quad (4)$$

$$CI = \frac{1}{N} \sum_{k=1}^K N_k (f_k - \bar{o}_k)^2 \quad (5)$$

In (3) to (5), $\bar{o}$ is the overall relative frequency of event occurrence (i.e., the base rate), K is the number of forecast categories (i.e., the number of ‘bins’ that probabilities are grouped into), N_k is the number of forecasts in category k , f_k is the probability assigned to forecasts in category k , and $\bar{o}_k$ is the relative frequency of event occurrence in category k .

A CI of zero indicates perfect calibration. Calibration is often displayed visually using calibration curves, where the x-axis shows the forecasted probability of an event occurring (or the subjective confidence of some proposition being true) and the y-axis shows the true proportion of times the event occurs (or the proposition is true) when the probabilities are binned into categories (Mandel & Barnes, 2014).

Perfect calibration is represented as a diagonal line on a graph, where there is perfect agreement between forecasted probability and the proportion of times an event occurs (see the black line in Figure 1.2). The yellow line in Figure 1.2 shows overconfidence, particularly at the higher end of the confidence axis. For example, when this forecaster declared an event would happen with 100% confidence (i.e., their subjective probability was 100%), the event occurred roughly 75% of the time, on average.

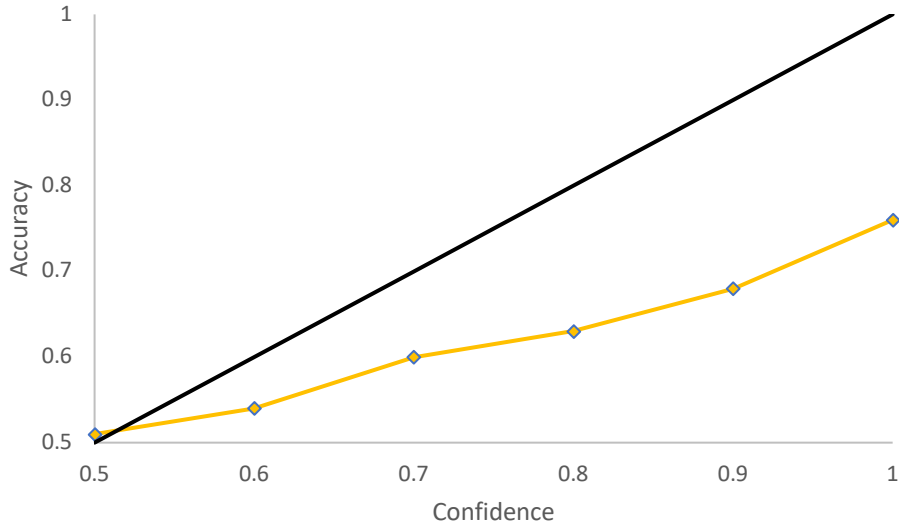


Figure 1.2. A calibration curve showing overconfidence

A different measure related to calibration is Bias (Yates, 1990). Bias measures the difference between a forecaster's overall average confidence and their proportion of correct predictions. Bias is given in (6), where $\bar{f}$ is the average confidence of a forecaster and $\bar{d}$ is their proportion correct.

$$Bias = \bar{f} - \bar{d} \quad (6)$$

A Bias of zero represents perfect calibration, whereas positive values represent overconfidence and negative numbers represent underconfidence. Bias is conceptually similar to the calibration index, except that Bias doesn't require multiple confidence categories to compute and thus requires less data to calculate a reliable measure.

One goal of the ACE Project was to find psychological variables that improve or predict forecaster accuracy (as measured by the Brier score), such as training in probabilistic and political reasoning, or working in teams (Chang et al., 2016). Another goal of the project was to develop algorithmic means of aggregating multiple forecasts to create a more accurate aggregate probabilistic predictions (Ungar et al., 2012).

In evaluating the impacts of the individual, situational, and other psychological variables, the researchers found that training in probabilistic and political reasoning, and working in teams significantly enhanced accuracy. Participants who engaged with the platform longer and more often, as measured by update

frequency and time spent deliberating before answering questions, also achieved greater accuracy (Mellers et al., 2015). Practice effects were additive over and above training, as participants who forecasted on a greater number of questions in total, or returned to the same question more often achieved greater accuracy, regardless of training (Chang et al., 2016). These effects are present for both the Brier score and its calibration component (Moore et al., 2017). This shows that calibration is a learnable skill that can be improved through deliberate practice. Individual differences in variables including IQ, Actively Open-Minded Thinking (AOMT), and scores on the Cognitive Reflection Test (CRT) also showed positive correlations with accuracy, indicating that beyond training and practice effects, there may be some inherent individual differences in forecasting ability (Mellers et al., 2015).

A separate review of predictors of forecasting accuracy found that of the non-accuracy variables related to forecasting skill, the most predictive were the average magnitude and frequency of updates to forecasts, followed by fluid IQ, then dispositional measures like conscientiousness, then political knowledge, education, and self-rated expertise (Atanasov & Himmelstein, 2023).

Another vein of research has investigated the extent to which calibration can be trained through deliberate practice with provision of different types of feedback. In their landmark study, Lichtenstein and Fischhoff (1980) trained participants to significantly improve the calibration of their confidence in general knowledge questions using multiple forms of feedback, including: the correct answers for all questions and the participant's answers to them, their proportion correct, overconfidence (Bias) score, the Brier score and its partitions, and calibration curves, which were all discussed in-depth by the experimenters after each training session. Notably, the training involved either 11 or 3 sessions (for experiment 1 and 2, respectively) of 200 questions per session. A similar study found that calibration feedback in the form of the calibration index and calibration curves improved the calibration of participants' predictions in a football forecasting task (Benson & Önköl, 1992). While these and other studies demonstrate that improving calibration through training is possible, particularly with performance feedback such as calibration curves, such training is often highly time-consuming and requires expert facilitation.

Other researchers have investigated whether calibration training can be automated and taught in relatively short training sessions. These studies often show mixed results. For example, Stone et al. (2023) found that their automated training paradigm had a positive effect on calibration for forecasts about the outcomes of baseball games, but not on a blackjack-related forecasting task. Gruetzemacher et al. (2023)

found that automated calibration training on general knowledge questions using a proper scoring rule had positive transfer effects to a football forecasting task, although the authors consider their findings to be a proof of concept requiring further evaluation. Their methodology also suffers from some shortcomings that are discussed more fully in Chapter 3.

Chapter 3 of this thesis evaluates a training paradigm designed to improve calibration through deliberate practice using a proper scoring rule intended to facilitate learning. It also seeks to examine individual differences like AOMT that might also predict ability to improve calibration.

1.6 Algorithmic Recalibration

Regarding improving the accuracy of forecasts using aggregation algorithms, the GJP discovered some techniques that could beat the accuracy of the unweighted mean of the probability forecasts of a particular question. The unweighted mean is an example of a linear opinion pool. An opinion pool is a method for combining multiple individual estimates of an unknown quantity (in this case, a probability) to form an aggregated opinion about that quantity (DeGroot & Mortera, 1991). A linear opinion pool is one where individual estimates are summed or averaged, often after applying a weight to the individual opinions (Satopää, Baron, et al., 2014). An ideal opinion pool maximizes both calibration and sharpness (Ranjan & Gneiting, 2010). Calibration, as discussed above, measures how close event frequencies are to forecast probabilities when binned into categories. Sharpness describes how far away the forecasts are from the marginal event frequency, or the naïve baseline forecast (e.g., .50 for a binary event given no information about event frequency). More extreme forecasts (i.e., closer to 0 or 1) are sharper than those closer to the baseline. The Brier score is affected by both calibration and sharpness and represents a summary measure of predictive ability (Ranjan & Gneiting, 2010).

Ranjan and Gneiting (2010) have demonstrated that while linear opinion pools are more accurate than individual forecasts, they are necessarily uncalibrated and lacking in sharpness. Linear opinion pools tend to produce aggregate forecasts that are too close to the marginal event frequency (e.g. pulled towards 0.5 when the baseline event probability is 0.5), leading to underconfidence. Non-linear recalibration of forecasts prior to linear pooling has been proven to improve aggregate accuracy (Ranjan & Gneiting, 2010). In this vein, Satopää et al. (2014) used a recalibration technique based on the logit function which was able to out-perform both the unweighted mean and the Beta-transformed linear opinion pool algorithm suggested by Ranjan and Gneiting (2010).

In follow-up work, [Satopää et al. \(2014\)](#) used a different aggregation technique better suited for dynamic data with sparse updates, like that of the GJP dataset, achieving more accurate results. Finally, [Atanasov et al. \(2017\)](#) used a combination of weighting by time, forecaster accuracy, updates per question, and recalibration to achieve aggregate forecasts with comparable or superior accuracy to the results of the previous studies outlined above.

A different approach to improving the accuracy of geopolitical forecasts is to use statistical or machine learning techniques applied to both the forecasted probabilities and other related predictor variables. To this end, [Shinitzky et al. \(2024\)](#) used Bernoulli naïve Bayes, gradient boosting classifier, linear discriminant analysis, and logistic regression to predict the outcomes of geopolitical events using a large number of predictors, including linguistic features derived from the forecast rationales and third-party evaluations of the convincingness of the rationales. They found that logistic regression was the most effective machine learning method. The most important predictor variables were forecast extremeness, difference of the forecast from the consensus mean forecast, and linguistic features of the rational related to time orientation.

While [Shinitzky et al.'s \(2024\)](#) results are interesting, there are some features of their experiment that motivate follow-up research. [Shinitzky et al. \(2024\)](#) used a relatively small sample of forecasters and questions compared to the GJP tournament dataset from the ACE competition. They also tested features that would be impractical to implement in a larger scale forecasting setting, such as third-party evaluations of forecast rationales. Furthermore, the authors did not incorporate more sophisticated forms of forecast aggregation, like those outlined above, which would likely have increased the overall predictive power of the models and would have made for an interesting comparison to the machine learning methods they tested. Chapter 4 of this thesis uses recalibration and aggregation in conjunction with BNs and logistic regression with the aim of increasing the overall predictive power of such modelling techniques. While [Shinitzky et al. \(2024\)](#) tested a Bernoulli naïve Bayes technique, BNs are able to represent complex conditional dependencies between predictor variables, which may increase their predictive power over naïve models. Bayesian networks also make fewer parametric assumptions about the dataset than logistic regression, which may prove advantageous with a larger and more varied dataset, such as from the GJP.

1.7 Discussion

This thesis develops and tests the effectiveness of cognitive and computational tools to help reduce bias and promote rational decision-making. The proposed tools encourage adherence to probability theory and other relevant epistemic and institutional norms. The main intended domain of application is the IC, given the importance of the decisions it makes and its relative past neglect of empirically validating its practices.

The body of the current thesis consists of three articles that contribute to the above aims. Chapter 2 empirically evaluates which analytic techniques are most predictive of high-quality reasoning and accuracy in a set of intelligence products. Some of the analytic tools under investigation are computational, such as BNs, whereas others are non-computational techniques, such as Timelines. The dataset under analysis allows for comparisons of the degree to which computational and non-computational tools relate to IC norm adherence and objective accuracy in an experimental context designed to replicate the conditions of the IC.

Chapter 3 consists of testing a training procedure for improving calibration of subjective probability judgements using a proper scoring rule called the Practical scoring rule (Greenberg, 2020). This rule is a modified version of the logarithmic scoring that is intended to facilitate learning calibration by being cognitively intuitive and easily interpretable. Discovering a new efficient means of training calibration would have practical benefits for organizations that use forecasting practices and would add to the current state of the literature on calibration training.

Chapter 4 consists of using BNs to predict the outcome of future geopolitical events from forecasts made during the GJP competition. It incorporates recalibration and aggregation as part of the predictor variables and contrasts the effectiveness of BNs to previously explored methods such as regularized logistic regression and recalibration and aggregation alone.

CHAPTER II

The Relationship Between Structured Analytic Techniques And Intelligence Analysis Quality: Evidence From Content Analysis³

2.1 Overview of Chapter 2

The literature review in Chapter 1 argued that the intelligence community (IC) has relied on structured analytic techniques (SATs) as a methodology to increase the rigor and transparency of analysis, despite relatively little empirical evidence demonstrating that the techniques are effective. For instance, Coulthart's (2017) systematic review found only 45 studies evaluating the efficacy of any of the 12 SATs in the *Tradecraft Primer* (US Government, 2009). In those 45 studies, only six of the 12 SATs were evaluated. Out of those six, only three had research studies of medium or high credibility, according to the Maryland Scientific Methods Scale (Farrington et al., 2002). These were analysis of competing hypotheses (ACH), brainstorming, and devil's advocacy. A separate review of ACH found that its effect on judgement and reasoning is likely not beneficial, and quite possibly net harmful (Wilcox & Mandel, 2024). A different review of scenario planning (a set of techniques which includes the SAT called alternative futures analysis) similarly finds a lack of empirical support in favour of its effectiveness (Cordova-Pozo & Rouwette, 2023). The literature review calls for more scientific work to empirically test the effectiveness of methods used by the IC to improve the rigor of their analyses.

The Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE) program was funded by Intelligence Advanced Research Projects Activity (IARPA) to develop tools to improve analytic rigor using SATs and crowdsourcing. One team in the program developed the Smartly-assembled Wiki-style Argument Marshalling (SWARM) platform, which gives analysts a wide degree of latitude in how they can approach problems given to them. The SWARM platform has been tested twice with teams of professional and amateur analysts who used SWARM to collaboratively respond to problems like those encountered in real IC analysis. The two tests yielded over 100 reports written by different teams of analysts in response to eight challenge problems. Reports were scored according to a version of the IC Rating Scale, which quantifies the degree to which analytic products adhere to the norms of analytic tradecraft (Office of the Director of National Intelligence, 2015b, 2015a). Additionally, five of the eight problems could be scored

³ Note to the reader: The journal this article was submitted to uses endnotes for citations and parenthetical notes. Parenthetical notes have been included as footnotes in this thesis.

for objective accuracy, two of which used applications of Bayes' theorem as the normatively correct answer.

The study in Chapter 2 conducts a content analysis of the reports created during the two tests of the SWARM platform. The study codes for the kind of analytic techniques used in each report. This allows for an evaluation of the relationships between the techniques used in reports, adherence to IC norms, and objective accuracy. This provides a rare empirical analysis of the extent to which using different analytic techniques predicts better adherence to IC norms of quality of reasoning, and accuracy.

2.2 The Relationship Between Structured Analytic Techniques and Intelligence Analysis Quality: Evidence from Content Analysis

Matthew Martin^{a*}

^a Department of Computer Science, Université du Québec à Montréal, Montréal, Canada

*Corresponding author

Email: martin.matthew@courrier.uqam.ca

LinkedIn: linkedin.com/in/mattmartin256

Orchid ID: 0009-0006-5765-9879

Publication status: Submitted to Intelligence and National Security on July 25, 2025.

Acknowledgements

I wish to thank Tim van Gelder, Ariel Kruger, and other members of the SWARM project team for the data, advice, and technical support that made this project possible. I also wish to thank my research assistants Bianca Abbandonato and Aimen Waqar for their hard work and invaluable contribution to this project. OpenAI's o3-deep-research tool was used to find relevant research on empirical evaluations of structured analytic techniques for the introduction section. Anthropic's Claude Sonnet 4 was used to suggest edits to the manuscript and improve language quality.

Article length: 10,150 words

2.3 Abstract

Structured Analytic Techniques (SATs) are intended to improve the rigour and transparency of intelligence analysis. This study conducted a content analysis of 138 reports from experiments simulating intelligence community (IC) conditions. The reports were scored for adherence to IC norms and accuracy, which correlated weakly. The techniques of Standardizing Probability Ranges and Timeline positively predicted IC norm adherence, as did the number of unique techniques in a report. No techniques reliably predicted accuracy, nor did the number of techniques used, likely due to low power. The results suggest that SATs have small positive correlations with IC norm adherence that compound.

Keywords: Structured Analytic Techniques, Intelligence Analysis, Content Analysis, Quality of Reasoning, Accuracy

2.4 Introduction

Improving intelligence community (IC) analysis represents a high-impact opportunity for applying decision science. Intelligence analysts must make sense of ambiguous, often conflicting information, weigh options, and suggest actions in matters of national security and foreign policy, often under considerable time pressure. In response to these cognitive challenges, practitioners have developed Structured Analytic Techniques (SATs) to help analysts structure information, mitigate biases, and make reasoning more transparent and logically sound (Chang et al., 2018). In the wake of perceived intelligence failures in the early 21st century, including the failure to anticipate the 9/11 attacks and the erroneous assessment that Iraq possessed weapons of mass destruction, the Intelligence Reform and Terrorism Prevention Act (2004) mandated the use of SATs in the US IC (Chang et al., 2018; Intelligence Reform Act, 2004). SATs also feature prominently in analyst training (Chang et al., 2018). Despite their strong institutional support, it appears that SATs are not routinely used within the IC and have weak evidence in support of their effectiveness in improving intelligence (Coulthart, 2016, 2017).

On the point of their limited use, a survey of 80 analysts in a US IC agency found that only 21% of analysts use them more than 'rarely'. The primary barriers to their use were a lack of training and reluctance to use techniques that have not demonstrated effectiveness at improving analysis (Coulthart, 2016). A separate review of unclassified IC reports showed that only 22 of the 63 examined documents (~35%) used SATs.

The authors commented that they found that most agencies do not systematically track the use of SATs in their products, nor do they evaluate how they affect quality of analysis (Artner et al., 2016).

The modest amount of research that has been done to investigate the effectiveness of SATs on improving analysis shows mixed results. For example, a recent systematic review of the 12 SATs advocated by the US government in 'A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis' found that only six had any empirical evaluation of their effectiveness: Devil's Advocacy, Analysis of Competing Hypotheses (ACH), Brainstorming, Alternative Futures Analysis, Red Teaming, and Team A / Team B (Coulthart, 2017; US Government, 2009). The proportion of studies supporting the effectiveness of those techniques ranged from 0 to 100%. Furthermore, only the first three techniques listed above were found to have evidence credibility higher than 'low' (Coulthart, 2017).

A more recent review of ACH, one of the most prominent SATs, showed that it provides little to no benefit on judgement quality and may even be detrimental. The critique argues that ACH suffers from lack of clarity in how to perform its key steps and that its procedure for evaluating the relative plausibility of competing hypotheses violates the norms of Bayesian reasoning (Wilcox & Mandel, 2024). Separately, a review on scenario planning (an umbrella category of techniques that includes the SAT Alternative Futures Analysis) found that there was a scarcity of research on the effectiveness of these techniques, along with conceptual confusion and 'methodological chaos' (Cordova-Pozo & Rouwette, 2023).

Mandel and Tetlock (2018) argue that the justification for using SATs has primarily been based on face validity – it appears that using SATs helps reduce bias, so encouraging analysts to use them probably does more good than harm (Mandel & Tetlock, 2018). The problem is that unaided human intuition is not a sufficiently reliable tool for assessing the effectiveness of cognitive interventions, where known biases such as bias blind spot and what-you-see-is-all-there-is (WYSIATI) are likely to occur (Mandel & Tetlock, 2018). Given the high-stakes nature of the judgements made by the IC, a higher standard of scientific rigor should be applied.

2.4.1 CREATE

While the IC does not routinely test the effectiveness of SATs, it sometimes funds high-risk research projects designed to improve its tradecraft through the Intelligence Advanced Research Projects Activity (IARPA). One such notable project was the Crowdsourcing Evidence, Argumentation, Thinking and

Evaluation (CREATE) program in which four projects tested novel ways of combining crowdsourcing and SATs to improve intelligence analysis (IARPA, 2016). The Smartly-assembled Wiki-style Argument Marshalling (SWARM) project from the CREATE program is the focus of the current study.⁴ The SWARM project used an online collaboration platform similar to Stack Exchange to cultivate engagement, collaboration and natural expertise among groups of analysts who would submit resources and draft reports that could be asynchronously rated and commented on (van Gelder et al., 2018).

The current study conducts a content analysis of the reports written during two evaluations of the SWARM platform. It assesses the relationship between quality of reasoning, accuracy, and the techniques used in the reports, including SATs. The SWARM project was chosen as the basis of the current study for a few reasons. First, the SWARM platform has been evaluated using teams of both trained intelligence analysts and members of the public, yielding many reports that can serve as a proxy for professional intelligence analysis. The 2018 Challenge and 2020 Challenge were designed to replicate the conditions of real intelligence analysis, with two main caveats. The first caveat is that participants generally worked on larger teams than is normal in the IC. While this may hamper generalizability, it likely increased the quality of reports and the number of techniques used, which is helpful for the current study. The second caveat is that some teams were composed entirely of members of the public, many of which had no training in intelligence analysis. This concern is tempered by the fact that, in aggregate, the public teams produced reports of comparable or better quality than the professional analyst teams, as discussed further below (Thorburn et al., 2020; van Gelder et al., 2020).

A second reason for selecting the SWARM results for secondary analysis is that it was unique in the CREATE program for how much flexibility it provided analysts in approaching the problems assigned to them and was thus more reflective of normal analytic conditions. Users were not constrained to use any specific SAT, but could approach problems as they saw fit, using the experience and judgement of their teammates to navigate the problems at hand. Users were provided with a set of resources in a living online encyclopaedia called the Lens Kit, which contained articles on concepts and techniques relevant to intelligence analysis, referred to as Logical Lenses. These articles included how to conduct various SATs, along with broader concepts and techniques such as how to use Bayes nets (van Gelder et al., 2018).

⁴ For details on the other CREATE projects, see: Boicu et al., (2018), Nicholson et al. (2019), and Stromer-Galley et al. (2021).

A third reason why the SWARM-generated reports were attractive is that they were evaluated similarly to real IC reports (Thorburn et al., 2022). Each report received a Quality of Reasoning (QoR) score measuring the extent to which reports adhered to the analytic tradecraft standards mandated in Intelligence Community Directive (ICD) 203 (see Methods for more detail; Office of the Director of National Intelligence, 2015). In addition to the intersubjective QoR score, five of the eight problems analysts completed could be assessed for objective accuracy, enabling an evaluation of how QoR and accuracy relate to each other and to the techniques found in the reports (Hunt Laboratory & University of Melbourne, 2020). This allows for a rare test of the validity of the IC rating system as well as the efficacy of SATs, as no ratings of reports conducted by the IC are publicly available (Thorburn et al., 2022). Given that one of the dimensions of the QoR score relates to accuracy, one would expect QoR and accuracy to be correlated. If SATs are effective at improving analytic quality, one would expect their use to be positively related to both QoR and accuracy.

The content analysis of the current study codes for the techniques and concepts used in the reports, using the Lens Kit as its main theoretical source of codes.⁵ The analysis first examines the correlation between QoR and accuracy. The degree to which the codes predict the QoR and accuracy of the reports is then assessed using Bayesian multiple regression analysis. The analysis also assesses whether a greater variety of codes in a report predicts QoR and accuracy. The first hypothesis is that QoR and accuracy are positively correlated. The second hypothesis is that at least one code is predictive of QoR and at least one predictive of accuracy. Based on the literature review above and given that the normatively correct answers to two of the five problems scored for accuracy are based on Bayes' theorem, one would expect techniques related to Bayes theorem to be more positively related to accuracy than ACH. The third hypothesis is that the presence of a greater number of unique codes in a report is positively related to both QoR and accuracy.

2.5 Methods

For the Participants and Materials subsections below, relevant details about the 2018 and 2020 Challenges are first recapitulated, followed by details pertinent to the current content analysis. The subsequent Procedure subsection only concerns the content analysis.

⁵ A note on the terminology used in this chapter: *code* refers to the use of one of the concepts defined in the coding manual and is the formal basis of the methodology and statistical tests. The codes are largely derived from the lens kit and the term *lens* is used roughly synonymously with code, although technically not all codes are lenses and vice versa. For variation of language, *technique* is also sometimes used roughly synonymously with code, even though not all codes are techniques and vice versa.

Participants provided informed written consent prior to participating. The Challenges were approved by the Human Research Ethics Committee of The University of Melbourne (project IDs 1646872.5 and 1646872.13). The secondary content analysis was approved by the Research Ethics Committee for Student Projects Involving Human Beings of l'Université du Québec à Montréal (certificate # 2023-5260). Data anonymization did not affect the results, as potentially identifiable participant characteristics were irrelevant to the study design.

2.5.1 Participants

2.5.1.1 SWARM Challenges

The 2018 Challenge recruited 641 participants split into 24 teams. For the current study, teams were categorized as Organizational teams if members were from intelligence-related government institutions, otherwise they were considered Public teams. While Public teams were recruited from the general population, many of them had considerable relevant experience with analysis. On average, Public teams were larger and spent more total hours on each report than Organizational teams (van Gelder et al., 2020).

In the 2020 Challenge, 20 teams of analysts completed reports in two stages with two problems each. For the first two problems, teams were either professionals from intelligence-related organizations or were from the public, similar to the 2018 Challenge. For the final two problems, the top performing members of the public teams were invited to form 'superteams', which professionals could also opt to join. For the current study, professional teams were categorized as Organizational teams, and public and super teams were considered Public teams. Public teams had more members and higher activity than Organizational teams, on average (Hunt Laboratory & University of Melbourne, 2020).

2.5.1.2 Content Analysis

The first author of this report and two hired coders conducted the content analysis of the reports generated during the 2018 and 2020 Challenges. Coders were required to have an advanced level of English reading comprehension⁶ and be able to learn to recognize statistical concepts including linear regression, probability theory, and Bayesian networks. Previous experience with content analysis was considered an asset but not required. All three coders had at least a master's degree from institutions in

⁶ Approximately level C1 of the Common European Framework of Reference for Languages (Council of Europe, 2025).

eastern Canada. The two hired coders were non-native speakers of English, one being of Québécois origin, the other South Asian, whereas the first author was a native English speaker of Canadian origin. All three coders had considerable experience with mathematics; the first hired coder was a college-level professor of mathematics and the other two had been teaching assistants for at least one university-level course in statistics. One coder was hired through connections at a Canadian university and the other was hired via social media.

2.5.2 Materials

2.5.2.1 Quality of Reasoning Score

All reports were evaluated using an adapted version of the Intelligence Community Rating Scale for Evaluating Analytic Tradecraft Standards (Office of the Director of National Intelligence, 2015b). This scale evaluates the extent to which an intelligence report adheres to the nine analytic tradecraft standards defined in ICD 203 (Office of the Director of National Intelligence, 2015a). The standards stipulate that an intelligence report:

1. Properly describes quality and credibility of underlying sources, data, and methodologies
2. Properly expresses and explains uncertainties associated with major analytic judgments
3. Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments
4. Incorporates analysis of alternatives
5. Demonstrates customer relevance and addresses implications
6. Uses clear and logical argumentation
7. Explains change to or consistency of analytic judgments
8. Makes accurate judgments and assessments
9. Incorporates effective visual information where appropriate

The rating scale defines each of the nine criteria in more detail and provides guidance on how to evaluate each of them as poor, fair, good, or excellent. For both the 2018 and 2020 Challenges, the seventh standard was not included as it was not relevant in the experimental setting.⁷ Reports were given a score from 1 to 4 on each of the eight remaining dimensions and the sum of the scores was used as the QoR score. Three

⁷ The rating rubrics for the 2018 and 2020 Challenges are available at: osf.io/cxsr8/

raters associated with the SWARM project independently evaluated each report, and the average of their ratings was used as the overall score for a report (Hunt Laboratory & University of Melbourne, 2020; van Gelder et al., 2020).

The QoR scoring technique is similar to the official implementation of the IC Rating Scale, where reports are independently rated by two evaluators, followed by a discussion and agreement on a consensus rating, followed by a third evaluator who assesses that evaluation for quality and makes any final changes. No aggregation procedure is prescribed for the official IC Rating Scale, however (Thorburn et al., 2022).

2.5.2.2 2018 Challenge

Participants in the 2018 Challenge used the SWARM platform to collaborate on writing reports in response to four fictional problems like those of the IC (labelled problems #1-4, see Table 2.1 for problem descriptions; van Gelder et al., 2020). Only problem #4 had an objectively verifiable answer, where a total of six probability estimates could be calculated using Bayes' theorem. The accuracy metric for this problem was the average absolute difference between the normative answers and the reports' probability estimates, subtracted from one. All accuracy metrics in the current study correspond to absolute error subtracted from one. Accuracy rather than error is used for consistency with past research as well as ease of interpretation (Hunt Laboratory & University of Melbourne, 2020).

Table 2.1. Problem Descriptions for the 2018 and 2020 Challenges

Problem	Challenge Year	Description	n_{Org}^a Reports	n_{Pub}^b Reports
1	2018	Determine the most likely causes of death of a politician	4	15
2	2018	Predict the optimal placement of assets for intercepting drug smugglers	2	17
3	2018	Assess the likelihood of suspects' involvement in terrorist activities and recommend countermeasures	4	16
4	2018	Calculate the probabilities of two nations acquiring weapons of mass destruction over time	4	15
5	2020	Geolocate four photographs and assess how they relate to two suspects' role in militant activities	7	13
6	2020	Estimate the probability of piracy and armed robbery against ships being greater than a certain frequency in a future month	7	13
7	2020	Calculate the probability an employee is stealing intellectual property for a rival company	6	5
8	2020	Evaluate four intelligence reports for reasoning errors and assess the connection between a suspect and a foreign intelligence agency	5	5
Total			39	99

^a n_{Org} = number of reports written by Organizational Teams

^b n_{Pub} = number of reports written by Public Teams.

2.5.2.3 2020 Challenge

Participants in the 2020 Challenge completed four problems (labelled #5-8). All four problems could be assessed for their objective accuracy in addition to QoR. In problem #5, participants geolocated four photos as part of their analysis; the proportion of correctly geolocated photographs was the accuracy measure. In problem #6, participants provided a probability forecast that the number of maritime criminal acts would be above some number in a future month. Accuracy on this question was the square root of the binary Brier score, subtracted from one. Problem #7 required participants to make three probability assessments that could be determined by Bayes' theorem. Accuracy was the average absolute difference from the normatively correct probabilities and the reports' estimates, subtracted from one. Problem #8 involved detecting four reasoning flaws that were intentionally put into fictional intelligence reports provided as part of the evidence base. Accuracy was the proportion of flaws correctly detected (Hunt Laboratory & University of Melbourne, 2020)

2.5.2.4 Content Analysis

For the content analysis, the unit of analysis in terms of data (i.e., 'sampling units') was a report written during either the 2018 or 2020 Challenge (Krippendorff, 2004). All available reports ($N = 138$) were coded to maximize statistical power and minimize selection bias.⁸ The unit of analysis for the coding scheme was a binary indication of whether a given code described in the coding manual was used at least once in the report under analysis.

The online Lens Kit formed the basis of most of the coding manual. In its form from the 2020 Challenge, the Lens Kit included over 75 entries ranging from tips on how to write clearly, to concepts such as regression to the mean, to how to use Bayesian network software. The final coding manual had 44 codes (also referred to as 'lenses') in eight categories mostly derived from the Lens Kit (see the Procedure section below for the development process).⁹

Each code in the coding manual contained a description, a set of inclusion and exclusion criteria for using the code, examples of uses of the code found in the reports, and negative examples where the code should not be applied (Syed & Nelson, 2015). Additionally, a set of tips and suggestions for applying the code was included for most entries. Codes with considerable conceptual overlap were formed into hierarchical groups with the rule that if a more specific member of the group applied, only it should be used and not a more generic member of the group. For example, the group related to numeric probabilities contained a generic code that applied to any use of numeric probabilities, a more specific code that applied to uses of Bayes' theorem, and more specific codes that pertained to specific applications of Bayes' theorem (such as Bayes nets).¹⁰ This hierarchical use of codes was intended to make the codes as mutually exclusive as possible and avoid multicollinearity in the eventual regression analyses.

The coding manual also contained instructions on how to use the corresponding coding sheet. The coding sheet was a spreadsheet where each row represented a report, and uses of codes were represented in

⁸ One 2018 Challenge report available for content analysis did not receive a QoR rating, most likely due to incompleteness, and it was also impossible to score for accuracy. It was thus excluded from analyses involving QoR and accuracy but was still coded as part of the content analysis.

⁹ A list of the lenses with a brief description of each is provided in Table SA1 and a redacted version of the coding manual is available in the online supplemental materials at: osf.io/8n6rz/

¹⁰ See Table SA1 and the coding manual in the supplemental materials for more information about the code hierarchy.

columns. Upon finding a codable lens in the report, the coder would select the code's name from a drop-down menu in a data-entry cell that contained all codes in the coding manual. The order of codes recorded in the columns did not matter and only one use of a lens within a report was coded. To aid the training process, the coding sheet was divided into four samples of 32 reports and a final sample of 10 reports. This was done to provide reliability feedback after a reasonable subset of reports had been coded, and to expose coders to a balance of all problem types at the beginning of training (see Procedure for the training process). To the extent possible, the samples were balanced by problem type and had four reports for each of the eight types, in random order. Because some problem types had less than 12 reports, only the first and second samples were perfectly balanced.

Krippendorff's α was used as the reliability metric in this study, computed using JASP (Jasp Team, 2023). This metric represents the degree to which coder agreement exceeds what would be expected by chance. It was used instead of Cohen's κ because κ is affected by irrelevant features of mismatches between coders, such as the distribution of which coder or category contributed more to the mismatches, making it unsuited for a reliability index for content analysis (Krippendorff, 2004). An $\alpha > .80$ was considered reliable, α between .667 and .80 was considered reliable enough for drawing tentative conclusions, and $\alpha < .667$ was considered unreliable (Krippendorff, 2004).

2.5.3 Procedure

The coding manual was developed through a 'theoretically driven inductive approach', where the Lens Kit and other SATs known to be important to the IC formed the theoretical background and primary source of codes (Syed & Nelson, 2015). Codes were added inductively to the coding manual when techniques appeared in the reports that were not already captured by the manual. Codes were also added to make the coding manual exhaustive and to cover generic cases, such as None when no lens was apparently used. The Other code was used as a catch-all for techniques that did not meet the description of any existing codes and was a source of new codes during the development process. Codes were consolidated or removed if they were rarely used, redundant with other codes, or were too vague to be coded reliably. The general rule was that codes that appeared less than three times should be removed or consolidated into a similar code, along with those that were irrelevant or too vague to be coded effectively. The exception to this rule was the 12 SATs of the 'Tradecraft Primer' (US Government, 2009). These were left in the manual despite low usage since they are important to the IC and their lack of use might itself be an

interesting finding. The coding manual was iteratively refined to make its application to the reports as clear and reliable as possible through the development process outlined below.

An initial set of 75 codes were developed by converting entries in the Lens Kit into codes. The coding manual was initially refined through an iterative process of the first coder coding the reports using the existing manual and making alterations to the codes after completing a sample. After a stable manual with 50 lenses across nine categories was reached after five iterations, all reports were re-coded using the fifth version of the manual. The second coder was then hired to help code and further refine the manual.

The second coder first read the coding manual and the problem descriptions, and any initial questions were addressed. The second coder then proceeded coding samples of 32 reports each. After coding each sample, reliability was assessed. Discrepancies between the two coders' codes were recorded and discussed and the coding manual was subsequently edited to improve clarity and reliability. This process was repeated four times, until the average α of the codes in a sample was $> .80$ and it was deemed that the coding manual was sufficiently clear. Once the final version of the coding manual with 44 codes was reached, the two coders re-coded all 138 reports using it.

Given that the second coder was involved in the development of the manual, they could not be considered an independent rater, as they were possibly influenced by information extraneous to the coding manual such as implicit agreements and conventions (Krippendorff, 2004). Thus, a third coder was hired to conduct a new independent coding. Due to feedback from the second coder, the third coder was trained iteratively on subsamples of eight reports balanced across problem type. After completing a subsample, discrepancies between the first and third coder's codes would be reviewed and discussed, but no changes to the manual were made. After six iterations of this training, the third coder was deemed to be sufficiently capable of using the coding manual independently. They then coded the remaining 90 reports with minimal interaction with the first and second coders. Coders coded roughly one to four reports per hour after training. The entire content analysis procedure took approximately 1500 person-hours of work, split across three people.

The initial 48 training reports of the third coder were not used in the reliability analysis described below as they do not reflect the work of a fully trained and independent coder. Disagreements in the 90 independently rated reports were resolved via majority vote. For the remaining 48 training reports, the first coder's analysis was considered the gold standard (Syed & Nelson, 2015).

2.6 Results

The reliability of the codes produced by the content analysis is examined first, followed by the tests of the three main hypotheses of the study. The first hypothesis that QoR is correlated with accuracy is tested first, followed by testing the relationships between QoR, accuracy, and codes using Bayesian regression analyses. Zero-order correlations between the dependent variables and codes are included to assist in the interpretation of the regression analyses. Finally, the third hypothesis that QoR and accuracy are predicted by the number of unique codes found in a report is tested with hierarchical regression analysis.

2.6.1 Reliability Analysis

The reliability of the 28 codes included in subsequent regression analyses (where $\alpha > .667$ and $n_{Full} > 1$) are reported in Table 2.2.¹¹ Most codes ($n = 28$) showed good reliability with $\alpha > .80$. Nine codes were reliable enough to only draw tentative conclusions. Seven codes had poor reliability with $\alpha < .667$. Low reliability appeared to be caused by either infrequent occurrences of the codes, making a single disagreement highly detrimental to α , or by confusion with conceptually or structurally similar codes.¹²

¹¹ n_{Full} = the number of times a code that appears in the full set of 138 reports. n_{Inde} = the number of times a code appears in the set of 90 reports coded by two fully trained independent raters. Reliability results for all 44 codes are in Table SA2 of the supplementary materials.

¹² See Supplementary Materials A for a fuller discussion of the causes of low reliability.

Table 2.2. Reliability Analysis of Codes

Code	n_{Full}^a	n_{Inde}^b	α	95% CI ^c
ACH	10	5	0.79	[0.32 - 1.00]
ADC	3	0	1.00	[1.00 - 1.00]
BN	9	7	1.00	[1.00 - 1.00]
BTH	15	8	0.93	[0.74 - 1.00]
BTR	3	1	1.00	[1.00 - 1.00]
CPTC	2	1	1.00	[1.00 - 1.00]
EVTB	12	8	0.81	[0.52 - 1.00]
GEOL	18	13	0.90	[0.73 - 1.00]
GEOM	52	32	0.95	[0.87 - 1.00]
GTPD	2	2	1.00	[1.00 - 1.00]
HYTR	4	4	0.74	[-0.01 - 1.00]
LINK	22	11	0.75	[0.44 - 0.95]
MMO	10	6	1.00	[1.00 - 1.00]
MTG	43	25	0.76	[0.59 - 0.90]
NONE	6	5	1.00	[1.00 - 1.00]
NPRG	102	66	0.79	[0.63 - 0.92]
ODNI	3	3	1.00	[1.00 - 1.00]
PRCN	6	4	1.00	[1.00 - 1.00]
RCIF	17	13	0.79	[0.58 - 0.95]
RGLM	10	6	0.71	[0.26 - 1.00]
RST	4	1	1.00	[1.00 - 1.00]
SPR	29	15	0.96	[0.85 - 1.00]
STCG	27	19	1.00	[1.00 - 1.00]
SWPC	11	8	0.85	[0.55 - 1.00]
TBLG	101	64	0.85	[0.71 - 0.95]
THPE	4	4	0.74	[-0.01 - 1.00]
TMLN	37	22	0.83	[0.69 - 0.94]
VSIM	25	17	0.71	[0.49 - 0.87]

Note. ACH = Analysis of Competing Hypotheses, ADC = Admiralty Code, BN = Bayes Nets, BTH = Bayes' Theorem, BTR = Bayes Tree, CPTC = Conditional Probability Table Calculator, EVTB = Evidence Table, GEOL = Geolocation, GEOM = Geographical Map, GTPD = Game Theory and Prisoner's Dilemma, HYTR = Hypothesis Tree, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques (Generic), NONE = None, NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, PRCN = Pros and Cons, RCIF = Reference Class-Informed Forecasting, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), THPE = Three-Point Estimating, TMLN = Timeline, VSIM = Visual Image (Other).

^a n_{Full} = number of times a code occurred all 138 reports

^b n_{Inde} = number of occurrences in the non-training subset of reports rated by two independent coders.

^c Krippendorff's α and its associated 95% confidence interval (CI) were calculated using 1000 bootstrap replications, where α represents the reliability between the two hired coders for the non-training reports.

2.6.2 Frequency of Code Usage

The distribution of the frequency of code usage was significantly right-skewed, with a mode of zero, median of 6.00 ($IQR = 16.25$) and mean of 14.82 ($SD = 22.49$). Eleven codes (25%) were used less than twice, most of which belonged to the 12 SATs described in the 'Tradecraft Primer'. Only two of these twelve SATs were used more than once: ACH and Key Assumptions Check. Numeric Probability and Odds (Generic) and Table (Generic) were the only codes used in most reports.

2.6.3 Quality of Reasoning and Accuracy

The hypothesis tests of this study used Pearson correlations and linear regression on the dependent variables of QoR and accuracy. To account for differences in problem difficulty and make the scores better fit the assumptions of linear regression analysis, both QoR and accuracy were standardized within a problem. These metrics thus reflect the score of a report relative to other reports on the same problem. Henceforth QoR and accuracy refer to their standardized forms.

Preliminary analyses were conducted to confirm the validity of correlation and linear regression models with QoR and accuracy as dependent variables. One of the main demographic differences between teams in the 2018 and 2020 Challenges was whether the members were recruited from the public or from government organizations with an intelligence function. Given that one of the aims of the study is to make inferences about how analytic techniques are used by members of the IC, significant differences in the behaviour of the two groups would indicate the model should adjust for team type.

An independent samples t -test showed that Public teams had significantly higher QoR ($M = 0.10$, $SD = 1.00$) than Organizational teams ($M = -0.26$, $SD = 0.89$) with a small effect size ($t(135) = 2.01$, $p = .047$, $d = 0.38$). Likewise, an independent samples t -test showed that Public teams had significantly higher accuracy ($M = 0.19$, $SD = 0.82$) than Organizational teams ($M = -0.33$, $SD = 1.14$) with a medium effect size ($t(78) = 2.37$, $p = .020$, $d = 0.55$). These differences are likely due to Public teams being generally larger, with more active participants and hours spent per report (Hunt Laboratory & University of Melbourne, 2020; van Gelder et al., 2020). To adjust for the impact of team type, the binary variable of Public Team was included as a control variable in subsequent analyses.

To test the first hypothesis that QoR and accuracy are positively correlated, a correlation analysis partialing out Public Team status was performed. Accuracy and QoR were significantly positively correlated with a

small effect size ($r(77) = .25, p = .028$)¹³. The zero-order correlation between QoR and accuracy was medium-sized ($r(78) = .30$). The first hypothesis was thus supported. Modest correlations between QoR and accuracy were similarly reported for the 2020 Challenge (Hunt Laboratory & University of Melbourne, 2020). This likely reflects the fact that making accurate judgments and assessments is only one of eight criteria that compose QoR, making the relationship between accuracy and QoR relatively small.

2.6.3.1 Regression Analyses

A few considerations established the choice of variables for inclusion in the regression models assessing the relationship between codes and QoR or accuracy. First, codes were excluded from the model if their reliability was $< .667$. The resulting data should be considered reliable enough to draw tentative conclusions, as the lowest α value in a multivariate dataset equals the joint reliability of the whole data (Krippendorff, 2004). Second, predictors were excluded if they were highly multicollinear with other predictors. The None code was excluded for this reason since it was perfectly predicted from the absence of other codes. Otherwise, an absolute correlation $r > .80$ or a variance inflation factor (VIF) > 5.0 was considered grounds for removal (no codes were removed for these reasons). Third, codes were excluded if they were used less than twice in the subset of reports under analysis. This represents the minimum number of positive cases required to estimate the standard error of a regression coefficient (β). While allowing codes with only two positive cases potentially increases the variance of β estimates and has other downsides discussed further below, the exploratory nature of the current study motivated including more variables and estimating their relationship with the dependent variables. After these preliminary cutoffs, 27 codes were suitable for the regression involving QoR and 22 for the regression involving accuracy.¹⁴

Ordinary least squares regression using this many predictors encounters problems such as data overfitting, unreliable β estimates, and poor generalization to other data sets (Myung, 2000). To ameliorate these problems while providing interpretable β estimates, this study used the Bayesian Adaptive Sampling (BAS) approach for multi-model inference with model averaging, as implemented in JASP (Bergh et al., 2021;

¹³ The p values reported in this chapter are not corrected for family-wise error. Given the three main hypothesis tests of this study (i.e., a partial correlation between QoR and accuracy, and a change in R^2 for both QoR and accuracy), a Bonferroni correction yields $\alpha = .0167$.

¹⁴ See Tables SA3 and SA5 in the supplementary materials for coefficient summaries for all included codes.

Clyde et al., 2011). The default JASP settings were used, with a JZS prior width = 0.125 and a beta binomial model prior where $\alpha = \beta = 1$.¹⁵

Intuitively, the BAS approach evaluates models using every possible subset of the predictors under consideration. For each model, an estimate of the relative probability of the model after observing the data is obtained (i.e., the posterior model probability) as well as estimates of the β s within each model. To estimate each β overall, a weighted average of their values across all models is taken, using the posterior model probabilities as weights. If the model space is too large to practically enumerate (i.e., the number of predictors is > 20), the model space is sampled efficiently without replacement to estimate models' posterior probabilities and β s (Bergh et al., 2021).

The first table for each analysis (see Tables 2.3, 2.6) contains a subset of the models (M) with the highest posterior probabilities $P(M/D)$. If an alternative model (M_1) has a higher posterior probability than the null model (M_0), it indicates that the predictors included in the alternative model can predict the data better than the null model. By default, the null model only contains the intercept as a predictor, but for the below analyses on QoR and accuracy, the null model also included the Public Team variable as a predictor. This means that the ability of codes to predict QoR or accuracy is after accounting for the predictive ability of team type, akin to hierarchical regression (Bergh et al., 2021).

The first table also includes the BF_{10} , which represents the degree to which the alternative model predicts the data relative to the null model. Following Bergh et al. (2021), BF_{10} s between 1 and 3 indicate anecdotal evidence in favour of the alternative model, BF_{10} s between 3 and 10 indicate moderate evidence, and BF_{10} s greater than 10 indicate strong evidence.

A secondary table includes estimates of the posterior distributions of the regression coefficients (β), including their model-averaged mean value, SD and 95% credible interval (see Table 2.4). The key metric indicating whether a predictor should be included in a model is the posterior inclusion probability, $P(incl/D)$, where a value $> .50$ indicates the predictor should be included in a regression model. The BF_{incl}

¹⁵ The robustness of the results was also tested using JZS prior widths of .25 and .50. The prior widths did not meaningfully affect the main findings.

quantifies the degree to which the data support including the predictor relative to excluding it and are interpreted similarly to BF_{105} .

2.6.3.2 Quality of Reasoning

The Bayesian regression analysis predicting QoR found that the most probable model included Standardizing Probability Ranges and Timeline as predictors. The second most probable model included only Standardizing Probability Ranges (see Table 2.3). Consistent with this, the posterior inclusion probabilities of model-averaged coefficients suggested that Standardizing Probability Ranges and Timeline should be included in a model. Their BF_{incls} indicated strong evidence in favour of their inclusion (see Table 2.4).¹⁶ The second hypothesis that at least one code would predict QoR was thus supported.

Table 2.3. Bayesian Regression Models Predicting Quality of Reasoning

Models	$P(M)$	$P(M D)$	BF_{10}	R^2
Null model (incl. Public Team)	0.161	0.007	1.00	0.03
SPR + TMLN	0.001	0.020	972.96	0.18
SPR	0.006	0.018	66.88	0.12

Note. All models include the Public Team variable, which is included as part of the null model. $P(M)$ = model prior probability, $P(M|D)$ = model posterior probability, BF_{10} = Bayes Factor of alternative model over null model, SPR = Standardizing Probability Ranges, TMLN = Timeline.

¹⁶ Table 2.4 includes a subset of coefficients most predictive of QoR. Table SA3 includes all coefficient summaries.

Table 2.4. Posterior Summaries of Coefficients Most Predictive of Quality of Reasoning

Coefficient	M_{β}	SD_{β}	95% Credible Interval		$P(incl)$	$P(incl/D)$	BF_{incl}
			Lower	Upper			
Intercept	0.00	0.08	-0.16	0.15	1.00	1.00	1.00
Public Team	0.27	0.16	-0.03	0.60	1.00	1.00	1.00
BN	0.16	0.28	-0.01	0.82	0.11	0.31	3.61
EVTB	0.25	0.33	0.00	0.91	0.12	0.46	6.15
HYTR	0.27	0.44	-0.01	1.25	0.12	0.35	4.13
SPR	0.33	0.27	0.00	0.76	0.14	0.71	15.51
TBLG	0.08	0.16	0.00	0.49	0.11	0.29	3.28
TMLN	0.23	0.22	0.00	0.61	0.13	0.62	10.73

Note. Public Team was included in the null model. M_{β} = mean β coefficient estimate, SD_{β} = standard deviation of β coefficient estimate, $P(incl)$ = prior inclusion probability, $P(incl/D)$ = posterior inclusion probability, BF_{incl} = Bayes Factor for inclusion, BN = Bayes Nets, EVTB = Evidence Table, HYTR = Hypothesis Tree, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), TMLN = Timeline.

2.6.3.3 Correlations

All zero-order correlations between QoR and the predictors were small ($\max |r| = .28$; Table 2.5).¹⁷ The highest zero-order correlations were with Timeline and Standardizing Probability Ranges. The results point to an absence of strong individual predictors of QoR. However, their impact may still be notable when combined, as examined by the tests of the third hypothesis further below.

¹⁷ Table 2.5 includes correlations where $|r| \geq .10$. Table SA4 includes all correlations with the predictor variables.

Table 2.5. Zero-Order Correlations with Quality of Reasoning

Variable	Pearson's <i>r</i>
Public Team	0.17
BN	0.18
EVTB	0.27
HYTR	0.22
LINK	0.17
MMO	0.15
MTG	0.16
NPRG	0.19
SPR	0.27
TBLG	0.24
THPE	0.12
TMLN	0.28

Note. BN = Bayes Nets, EVTB = Evidence Table, HYTR = Hypothesis Tree, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques (Generic), NPRG = Numeric Probabilities and Odds (Generic), SPR = Standardizing Probability Ranges, TBLG = Table (Generic), THPE = Three-Point Estimating, TMLN = Timeline.

2.6.4 Accuracy

None of the models predicting accuracy had a higher posterior probability than the null model. The most likely of the non-null models was the model that only included Timeline as a code, with the BF_{10} suggesting moderate evidence in favour of the model (see Table 2.6). For the model-averaged regression coefficients, all codes had a posterior probability of inclusion $< .5$, suggesting insufficient evidence to include any of them in a regression model.¹⁸ All BF_{incls} suggested only anecdotal evidence in favour of inclusion (or exclusion) of any parameter. The second hypothesis that at least one code would be predictive of accuracy was thus not supported.

¹⁸ See Table SA5 for summaries of all coefficients predicting accuracy.

Table 2.6. Bayesian Regression Models Predicting Accuracy

Models	$P(M)$	$P(M/D)$	BF_{10}	R^2
Null model (incl. Public Team)	0.116	0.123	1.00	0.07
TMLN	0.005	0.019	3.43	0.13

Note. All models include the Public Team variable, which is included as part of the null model. $P(M)$ = model prior probability, $P(M/D)$ = model posterior probability, BF_{10} = Bayes Factor of alternative model over null model, TMLN = Timeline.

The lack of predictive ability of the coefficients was likely owing to the low sample size of reports that could be coded for accuracy ($N = 80$) and high complexity of the model. A post-hoc power analysis of a regression analogous to the above indicated that a sample size of $N = 126$ would be required to detect a statistically significant single regression coefficient with a partial correlation as strong as the largest individual predictor.¹⁹

2.6.4.1 Correlations

The zero-order correlations between accuracy and its predictors were all small, with the highest correlation being with Timeline ($r = .28$; see Table 2.7).²⁰ The zero-order correlations for codes related to Bayes theorem (Bayes' Theorem and Bayes Nets) were more positive than for ACH, which was slightly negative ($r = -0.08$). There was thus suggestive, but inconclusive, evidence that ACH was more negatively associated with accuracy than Bayesian methods, in line with the prediction about it. Overall, the results point to an absence of strong individual predictors of accuracy, but do not rule out the possibility of combined effects.

¹⁹ The details of the power analysis are available in Supplementary Materials A.

²⁰ Table 7 contains a subset of correlations with accuracy. Table SA6 contains all correlations.

Table 2.7. Zero-Order Correlations with Accuracy

Variable	Pearson's <i>r</i>
Public Team	0.26
ACH	-0.08
ADC	0.10
BN	0.14
BTH	0.22
EVTB	0.19
GEOM	0.10
LINK	0.14
MMO	0.15
NPRG	0.15
ODNI	-0.19
RGLM	0.10
RST	0.19
TMLN	0.28
VSIM	0.15

Note. ADC = Admiralty Code, BN = Bayes Nets, BTH = Bayes' Theorem, EVTB = Evidence Table, GEOM = Geographical Map, LINK = Link Analysis, MMO = Means Motive Opportunity, NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, TMLN = Timeline, VSIM = Visual Image (Other)

2.6.5 Diversity of Codes

To test the third hypothesis that the number of unique codes used in a report predicted QoR and accuracy, hierarchical linear regressions were conducted on these dependent variables with two blocks of predictor variables. The first block included Public Team as a predictor and the second block included number of codes in a report as a predictor. A significant change in R^2 from the first to second blocks would indicate that the number of unique codes was related to the dependent variable, after accounting for the effect of Public Team.

For the regression with QoR as the dependent variable, the results showed that the second block showed a significant improvement over the first block with a medium effect size ($\Delta F(1,134) = 33.29, p < .001, \Delta R^2 = .19$). The regression coefficient for number of codes was significantly positive ($b = .16, t = 5.77, p < .001$) and the zero-order correlation was medium-sized, $r(135) = .46$. The results thus supported the third hypothesis that the number of unique codes in a report predicts QoR.

For the hierarchical regression with accuracy as the dependent variable, the results showed that the second block did not show a significant improvement from the first block and the effect size was small, $\Delta F(1,77) = 2.14, p = .15, \Delta R^2 = .03$. The regression coefficient for number of codes was non-significantly positive ($b = 0.07, t = 1.46, p = .15$) and the zero-order correlation was small, $r(78) = .23$. The results thus did not support the third hypothesis that the number of unique codes in a report predicts accuracy.

2.7 Discussion

This study explored the relationships between the kinds of techniques used in intelligence reports and the QoR and accuracy of those reports. It employed a content analysis methodology to code the techniques that were used in reports written during two experiments that attempted to replicate the conditions of the IC. The hypotheses under consideration were, first, that QoR and accuracy would be positively related, second, that some coded techniques would predict QoR and accuracy, and third, that the number of unique codes in a report would predict QoR and accuracy. The first hypothesis was supported and the second and third hypotheses were supported for QoR, but not accuracy.

Prior to testing the hypotheses using correlation and regression techniques, the reliability of the codes was examined. Most codes were found to be reliable enough to draw at least tentative conclusions. Many codes were used rarely, including the 12 techniques advocated in the 'Tradecraft Primer', of which only ACH and Key Assumptions Check were used more than once (US Government, 2009). The rarity of these codes in the dataset is likely partly contextual - the techniques were not used since they were not very relevant to the problems analysts addressed. Nevertheless, the lack of high-profile techniques in the dataset highlights the need for a partially bottom-up approach to content analysis for this kind of data, where the coding manual is adapted to describe the data at hand rather than rely on purely theoretical or prescriptive norms about which techniques should be used.

The first hypothesis that QoR and accuracy are positively correlated was supported, although the effect size was small. While the small effect might be surprising at first glance, it is important to consider that the QoR score is a simple addition of eight measures of adherence to norms important to the IC, only one of which is accuracy. While it could be argued that this provides a more wholistic assessment of quality, some researchers have questioned the validity of the QoR score for measuring the constructs of high-quality reasoning or analytic rigor, since reports can be considered high quality while relatively inaccurate.

They advocate for a more sophisticated aggregation procedure or using different measures of analytic rigor for these reasons (Marcoci et al., 2019; Thorburn et al., 2022).

The second hypothesis that some coded techniques would predict QoR scores was supported. The Bayesian linear regression found that the codes Standardizing Probability Ranges and Timeline were positively related to QoR. A few explanations for why these two codes were found to be related to QoR present themselves. First, these codes were among the most frequently used, so there would be less uncertainty in the model regarding their relationships with QoR. Relatedly, these two codes, along with the other frequently used codes, tended to be positively correlated with the total number of codes used within a report, another known correlate of QoR. The other frequently used codes tended to be generic codes that were widely present in both high-quality and low-quality reports. For example, Numeric Probability and Odds (Generic) could be applied to weather reports (e.g., 80% chance of rain) and demographic information interpretable as conditional probabilities (e.g., 90% of terrorists are male), so the presence of this lens in a report was not very indicative of its quality. In contrast, Standardizing Probability Ranges and Timeline appear to be not so ubiquitous such that their inclusion in a report was not akin to noise.

Standardizing Probability Ranges is also straightforwardly related to the criterion of ‘Properly expresses and explains uncertainties’. The scoring rubric from the 2018 Challenge explicitly states that such standardized probability ranges must be used to express probability (in line with ICD 203), and the 2020 Challenge scoring rubric likewise requires consistency in the terms used to describe probability (Office of the Director of National Intelligence, 2015a). Including such probability ranges is thus a strong indicator of adherence to IC norms.

There are fewer apparent reasons why Timeline was related to QoR. Timelines could take the form of line-type diagrams, tables, or ordered lists (corresponding to the SAT of Chronologies and Timelines from Pherson and Heuer (2021)). Timeline appeared in reports of all question types except for #2, which required participants to predict the activities of drug smugglers and recommend the optimal choice of assets to intercept them. Timeline is most clearly related to the criterion of ‘Incorporates effective visual information’. The use of tables as a means of visually organizing information may have been generally beneficial for this criterion, since other table-related lenses such as Table (Generic) and Evidence Table also had relatively high zero-order correlations with QoR ($r = .24, .27$, respectively). Qualitatively, Timeline

appears to have been used to summarize information, discover patterns, assess hypotheses for discrepancies, and calculate probabilities (in the case of Bayesian-type questions). It was thus also likely beneficial in creating a report with 'Clear and logical argumentation'.

Pherson and Heuer (2021) consider Chronologies and Timelines to be a diagnostic SAT useful in revealing important patterns, highlighting gaps in information, and reducing cognitive biases. Little empirical work has been done to support their specific claims, although some studies provide suggestive evidence of the general benefits of timelines. A timeline tool was part of the TRACE system developed under the CREATE program, along with other tools inspired by other classic SATs. The researchers found that higher quality analytical products were written by users who used the timeline tool earlier in the analytic process rather than later (Stromer-Galley et al., 2021). Research from the domain of digital forensics suggests that activity timeline tools are important in understanding incidents under investigation, but that some visualization tools may produce tool-induced confirmation bias due to misleading presentation of information (Andersen et al., 2024). Together these findings suggest that Timelines are a good tool for establishing appropriate context early in an analysis which may help prevent biases such as confirmation bias, which in turn helps analysts judge the likelihood of competing hypotheses with a more complete set of information. More research investigating the conditions under which timelines reduce specific cognitive biases should be done to better understand how the technique can be best used in intelligence analysis.

The Bayesian regression analysis predicting accuracy from coded techniques found that no codes were conclusively related to accuracy, although Timeline had modest evidence in its favour. Given that Bayes' theorem is a normative account of reasoning under uncertainty based in probability theory, and that two of the problems coded for accuracy relied on it, it was expected that Bayesian-related codes would be predictive of higher accuracy, especially in contrast to ACH, which violates the norms of Bayesian reasoning, among other shortcomings (Wilcox & Mandel, 2024). The more positive zero-order correlations of Bayes' Theorem and Bayes Nets with accuracy compared to ACH lend some credence to this hypothesis, but firm conclusions cannot be drawn due to lack of statistical power. While Bayesian techniques were primarily used to solve problems where Bayes' Theorem provided the normative answer, they were used in other contexts as well. More research should examine how Bayesian techniques can be better incorporated in intelligence analysis, as they provide a normative account of probabilistic reasoning and a number of studies have shown them to be effective in contexts relevant to the IC (Korb et al., 2020; Mandel & Navarrete, 2015).

The third hypothesis that the number of unique codes in a report would predict QoR and accuracy was supported for QoR but not accuracy. The positive relationship between the number of unique codes and QoR has a few explanations. Incorporating more techniques and concepts in a report is likely correlated with the amount of time spent crafting a report. The QoR score was correlated with metrics of participant activity and engagement in both the 2018 and 2020 Challenges (Hunt Laboratory & University of Melbourne, 2020; van Gelder et al., 2020). A plausible hypothesis is that more engaged members spent more time applying different lenses and adding to the overall quality of reports in other ways. Applying multiple analytic frameworks is also related to the criterion of ‘Incorporates analysis of alternatives’, which assesses the degree to which analysts evaluate different hypotheses, explore possible outcomes, and imagine possible futures. The QoR score is also multifactorial, so incorporating multiple types of techniques likely makes a report score higher on individual sub-scores (e.g., effectively incorporating visual information and properly expressing uncertainties likely involve using more than one technique).

The smaller and non-significant relationship between the number of unique codes and accuracy was unexpected. The constrained nature of the problems where accuracy could be calculated in this study may play a role. High accuracy could usually be achieved by the proper application of a single technique in each of the problems (e.g., Bayes’ Theorem for problems #4 and #7), whereas application of multiple techniques inappropriate to the task could be detrimental. The content analysis coded for the use of a technique, but not whether it was appropriate or implemented well. This was done to avoid circularity with QoR, which already captured the quality of a report. This said, prior research on the efficacy of SATs suggests they are not all equally effective and indeed some of them may be detrimental to accuracy (Coulthart, 2017; Wilcox & Mandel, 2024). The results of the current study further highlight the need for more research on the relationship between SAT use and accuracy and argue against relying on face validity or the intuitive appeal of SATs to support their use.

2.7.1 Limitations and Future Directions

This study had some methodological constraints and data-related issues that likely contributed to finding relatively few reliable predictors of QoR and accuracy. Regarding methodological constraints, the partially bottom-up approach to the content analysis yielded a high number of codes to record. While this helped capture the diversity of codes in the reports, the high number of codes in the analyses increased the complexity of the regression models, reducing their power. The fixed and somewhat small sample size of reports also reduced power and introduced pressure to complete the training portion of the content

analysis using fewer reports, which likely had a negative impact on reliability. This said, including a greater variety of codes was warranted given the exploratory nature of the study which sought to describe the character of the data rather than test constrained hypotheses. The fact that Timeline was the technique most associated with QoR and accuracy was a surprise. Given the relative lack of research on the effectiveness the techniques used by the IC in naturalistic contexts, more surprises like this should be expected, and exploratory work should be done in tandem with controlled experiments to both generate new hypotheses and test them rigorously.

Regarding data issues, the zero-order correlations between the predictor and dependent variables were all small, again lowering statistical power. This is likely a result of the observational nature of the study, where beneficial techniques were present in reports of both high and low quality. The lack of robust correlation between, for example, Geographical Maps and accuracy does not imply using maps is invalid, but rather that their benefit could not be distinguished from noise in this dataset, possibly owing to their ubiquity. The small effect sizes of all predictors suggest that no single technique was especially predictive of the QoR and accuracy of reports, which is itself an important finding. Future researchers should take this into account when considering the power of their studies.

Questions also remain about the generalizability of the current findings to real IC contexts. While the 2018 and 2020 Challenges were designed to replicate the conditions of the IC and included teams of professional intelligence analysts, some of the problems were somewhat artificial, particularly those that required calculating probabilities using Bayes' Theorem, which would be rarely used in real work (Thorburn et al., 2022). The stakes and time pressures of the voluntary experiment were also not the same those in the IC, where important decisions must often be made in a few hours (van Gelder et al., 2020). This said, the controlled nature of the experiments had some advantages. Multiple teams completed the same problems under similar circumstances, making evaluations between reports within problems more meaningful and allowed for controlling for problem difficulty. A quasi-replication of this study using unclassified reports from the IC would allow for a more realistic examination of how technique use relates to QoR and accuracy.

One further consideration for future studies is that the high number of person-hours required for manual content analysis could be significantly reduced using artificial intelligence. Large language models have been shown to be effective at content analysis (Chew et al., 2023; Pilny et al., 2024). A future study could use these models to rate a greater number of unclassified reports for QoR and code for the techniques

used in them. Reliability could be confirmed with a manual review of a subsample of the reports. The results of the current study could be used as a basis for these future projects as a potential training or validation dataset.

2.7.2 Conclusions

This study was one of relatively few examining the relationships between accuracy, quality of reasoning, and the analytic techniques used in intelligence reports. Quality of reasoning and accuracy were modestly positively related. The techniques of Standardizing Probability Ranges and Timeline were found to predict quality of reasoning, as did the total number of techniques used in a report. There were no reliable predictors of accuracy, likely owing to insufficient statistical power caused by small effect sizes and a smaller sample of reports. The study highlights the need for more research examining these relationships in naturalistic contexts using larger samples, potentially using automated tools.

2.8 Funding details

This work was supported by the Fonds de recherche du Québec – Nature et technologies under a doctoral training scholarship to Matthew Martin, Grant #303239.

2.9 Disclosure statement

The author reports there are no competing interests to declare.

2.10 Data availability statement

The author confirms that the data supporting the findings of this study are available within the supplementary materials found at: osf.io/8n6rz/.

CHAPTER III

Calibration Feedback with the Practical Scoring Rule Does Not Improve Calibration of Confidence²¹

3.1 Overview of Chapter 3

In Chapter 2, an observational methodology was used to examine the relationships between the techniques used in intelligence analysis (including structured analytic techniques (SATs)) and measures of quality of reasoning and accuracy. The chapter provides some clues as to which techniques may be more beneficial for accurate and high-quality reasoning, such as the Timelines technique. However, due to factors like the observational design, and the relatively small sample size, inferences about causality cannot be made, and the correlations between technique use and quality of reasoning and accuracy were small. Experimental designs are needed to determine which cognitive tools or techniques play a causal role in improving reasoning or judgement, allowing stronger claims than Chapter 2 could make.

Analysis tools like SATs are designed to externalize the thinking processes behind judgements and provide a framework for analysis that is intended to reduce bias and make reasoning more logical and transparent. However, as outlined in the literature review in Chapter 1, the imposition of structured frameworks may in fact increase the bias and noise of judgements in some cases. This is arguably the case with analysis of competing hypotheses (ACH), which requires analysts to use multiple ambiguous steps, and ignore confirmatory evidence in violation of the norms of Bayesian reasoning (Chang et al., 2018).

Rather than using external frameworks to debias judgements, an alternative approach is to directly train analysts' intuitions to be less biased. Calibration training using proper scoring rules that disincentivize both under- and overconfidence in probabilistic judgements are one way to reduce bias. Proper scoring rules are unlikely to lead to increases in bias or noise due to their basis in probability and decision theory. It is theoretically proven that the best way to maximize a proper score is to honestly report one's true subjective probability distribution and thus minimize miscalibration (Gneiting & Raftery, 2007).

²¹ Note: The online Supplementary Materials for this article have been reproduced in Supplementary Materials B of this Thesis

The Good Judgement Project (GJP) examined the role of both training and practice on the accuracy of forecasters in their multi-year forecasting tournament (as measured by the Brier score). They found that a short, one-hour training session on probabilistic reasoning techniques such as using comparison classes and computing a base rate was successful at increasing accuracy. On top of the effect of this training, practice also showed a positive relationship with accuracy, where forecasters who provided more forecasts per question or more total forecasts showed greater accuracy than those with fewer (Chang et al., 2016). Forecasters in the GJP tournament received outcome feedback (i.e., the outcome of the event being forecasted) and performance feedback (in the form of the Brier score (Ungar et al., 2012)).

The practice effects reported above took place over the course of months of forecasting. While this kind of long-term practice on similar tasks is likely the most robust, there may be other means of increasing accuracy with deliberate practice that do not require as long to achieve. Previous studies have shown that training calibration of subjective confidence with long-run accuracy on general knowledge questions is possible using performance feedback such as the calibration component of the Brier score, or calibration curves (e.g., Benson & Önköl, 1992; Lichtenstein & Fischhoff, 1980). Some studies have shown transfer effects from calibration training on general knowledge questions to forecasting tasks (Gruetzemacher et al., 2023; Stone et al., 2023). However, most successful calibration training experiments have required many hours of training, often with specialized participants or trained facilitators (Benson & Önköl, 1992; Lichtenstein & Fischhoff, 1980; Stone & Opel, 2000). The study in Chapter 3 investigates whether calibration can be improved through a short training session by using the Practical scoring rule as the primary performance feedback metric. The Practical scoring rule is designed to be intuitive to participants and can be provided on a trial-by-trial basis (unlike calibration curves) and is thus a candidate for successful short-term calibration training. If successful, this kind of training could be incorporated in forecasting training paradigms. Chapter 3 is adapted from the article by Martin and Mandel (2025).

3.2 Calibration Feedback With the Practical Scoring Rule Does Not Improve Calibration of Confidence

Matthew Martin¹ | David R. Mandel²

¹Department of Computer Science, Université du Québec à Montréal, Montréal, Québec, Canada |

²Intelligence, Influence and Collaboration Section, Toronto Research Centre, Defence Research and Development Canada, Toronto, Ontario, Canada

Correspondence: David R. Mandel (david.mandel@forces.gc.ca)

Received: 30 May 2024 | Revised: 25 September 2024 | Accepted: 15 October 2024

Funding: This research was supported by a doctoral training scholarship to the first author by the Fonds de recherche du Québec—Nature et technologies (#303239). It was also supported by the Accelerate Command, Control, and Intelligence Project Activities #AC2I - 019 and AC2I - 031 and Canadian Safety and Security Program project #2018 - TI - 2394 under the direction of the second author.

Keywords: calibration | feedback | individual differences | scoring rules | training

3.3 Citation

Martin, M., & Mandel, D. R. (2025). Calibration Feedback With the Practical Scoring Rule Does Not Improve Calibration of Confidence. *Futures & Foresight Science*, 7(1), e199. <https://doi.org/10.1002/ffo2.199>

3.4 Copyright

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 His Majesty the King in Right of Canada. Futures & Foresight Science published by John Wiley & Sons Ltd.

3.5 Abstract

People are often overconfident in their probabilistic judgments of future events or the state of their own knowledge. Some training methods have proven effective at reducing bias, but these usually involve intensive training sessions with experienced facilitators. This is not conducive to a scalable and domain - general training program for improving calibration. In two experiments ($N_1 = 610$, $N_2 = 871$), we examined the effectiveness of a performance feedback calibration training paradigm based on the Practical scoring rule, a modification of the logarithmic scoring rule designed to be more intuitive to facilitate learning. We examined this training regime in comparison to a control group and an outcome feedback group. Participants were tasked with selecting which of two world urban agglomerations had a higher population and to provide their confidence level. The outcome feedback group received

information about the correctness of their choice on a trial - by - trial basis as well as a summary of their percent correct after each experimental block. The performance feedback group received this information plus the Practical score on a trial - by - trial basis and information about their overall over - or underconfidence at the end of each block. We also examined whether Actively Open - Minded Thinking (AOMT) was predictive of calibration and its change across blocks. We found no improvement in calibration due to either training regime. Good calibration overall was predicted by AOMT, but not its change across blocks. The results shed light on the generalizability of other findings showing positive effects of performance training using the Practical scoring rule.

3.6 Introduction

When faced with making decisions in the face of incomplete knowledge, people must make probabilistic judgments, either implicitly or explicitly. Whether predicting the outcomes of future events (e.g., will this project be completed on time?) or assessing the likelihood of unknown current possibilities (e.g., are the terrorists hiding in this house?), probabilistic judgments must be made by both individuals and organizations. Unfortunately, human judgment is biased in many ways, with overconfidence being a common and significant impediment to accuracy (Lawrence et al., 2006; Moore & Schatz, 2017). Overconfidence in the context of probabilistic judgments about discrete events is a form of miscalibration, where subjective confidence is higher than long - run accuracy would warrant. A perfectly calibrated judge would be correct 80% of the time whenever they are 80% confident, whereas an overconfident judge would be correct less often and an underconfident judge would be correct more often. While less commonly observed, underconfidence has been documented in some expert judgments, such as by intelligence analysts making geopolitical forecasts (Mandel & Barnes, 2018). Underconfidence may be less troubling than its polar opposite because it implies the judge is correct more often than they believe rather than more incorrect than they believe, as in the case of overconfidence. Nevertheless, underconfidence can water down the informativeness of probabilistic assessment for decision - makers and more broadly undermine the quality of information processing and decision - making (Biais et al., 2005; Mandel & Barnes, 2014; Zacharakis & Shepherd, 2001). Accordingly, judges should strive to be well - calibrated.

Miscalibration is a widespread phenomenon that has been observed in both experts and nonexperts (Arkes, 2001; Bier, 2004), ranging in disciplines from medicine (Benjamin et al., 2022; Berner & Graber, 2008; Meyer et al., 2013), finance (Ben-David et al., 2013; Biais et al., 2005; Du & Budescu, 2007),

education (Callender et al., 2016; Foster et al., 2017; Huff & Nietfeld, 2009), sports betting (Bum et al., 2018; Erceg & Galić, 2014), the military (Phelps et al., 1980), and the intelligence community (Mandel & Barnes, 2018). The consequences of miscalibration include systematic diagnostic errors in medicine (Berner & Graber, 2008), financial losses due to overoptimism and misperception of risk in corporate finance (Ben-David et al., 2013), and failure to anticipate important geopolitical events relevant to governments and militaries (Tetlock, 2006). Successfully training people to become better calibrated—namely, reducing both overconfidence and underconfidence—would thus have benefits from several domains ranging from medical and financial to political and military.

Some calibration training methods appear to be successful at reducing bias, although these methods have inconsistent results and other drawbacks. Deliberate practice using performance feedback (such as showing participants calibration curves after a series of trials) can be an effective means of improving calibration. However, the most successful historical attempts required hundreds of practice trials and intensive feedback sessions with participants (Benson & Önköl, 1992; Lichtenstein & Fischhoff, 1980; Stone & Opel, 2000). Performance feedback in the form of the Brier score over the course of a multi-year forecasting tournament was also shown to be beneficial to calibration, as were dispositional variables like Actively Open - Minded Thinking (AOMT; a measure of open-mindedness) and general intelligence (Mellers et al., 2015).

More recently, Stone, et al. (2023) found that an automated form of individualized feedback using calibration curves and detailed textual advice could reduce overconfidence in a forecasting task about whether players would bust in a modified version of blackjack. However, the automated feedback regime was not effective at reducing overconfidence in a more realistic forecasting task of predicting baseball games, so more work is needed to determine the conditions under which such feedback is effective. Other studies have shown positive results for more simple kinds of training such as providing outcome feedback, where participants are simply informed about the outcome of each prediction (Du et al., 2012; Subbotin, 1996). However, outcome feedback is usually found to be less effective than performance feedback (Lawrence et al., 2006).

Other researchers have shown that training sessions that teach participants best practices about probabilistic forecasting can be effective at improving calibration in a business context (Hernandez, 2017) and in forecasting tournaments (Mellers et al., 2015). However, these sessions were highly

context-specific, requiring either expert facilitators or a subject pool of highly motivated forecasting enthusiasts. In this vein, a recent study on the effectiveness of a multi - hour, expert - facilitated commercial calibration training service on intelligence analysts found it was effective at improving intelligence analysts' calibration of 90% confidence interval estimates, but detrimental for point estimates of confidence regarding binary events, where underconfidence was exacerbated (Kelly & Mandel, 2024). This finding is consistent with a more general point that many interventions to reduce bias focus on reducing overconfidence, with a potential backfire effect of exacerbating underconfidence (Chang et al., 2018). Interventions that account for the bipolar nature of this kind of bias, such as calibration training using proper scoring rules, are an attractive approach to reducing error in both directions. In sum, a context - general and robustly effective method for training good calibration that does not require expert facilitation, specialized participants, or hours of intensive training would both advance the state of the research literature and provide a practical tool that individuals and organizations could use to improve accuracy in judgment.

3.6.1 Practical Scoring Rule

With the aim of developing an effective and scalable calibration training method, Greenberg (2020) developed the Practical scoring rule intended for use in an online calibration training app. The Practical scoring rule is a proper scoring rule for categorical judgments that is a modification of the logarithmic scoring rule intended to ameliorate some of its unintuitive features that would impair learning in a training setting. As a proper scoring rule, the logarithmic rule incentivizes forecasters to give honest predictions of how likely they believe a hypothesis is true. However, the logarithmic scoring rule disincentivizes both overconfidence and underconfidence in a somewhat counterintuitive way. By default, the logarithmic score has a maximum of zero (for perfect calibration) and a minimum of negative infinity (when a forecaster gives a prediction with 100% confidence and is wrong). The unipolarity of the logarithmic scoring rule (with zero being a maximum) makes it difficult to distinguish between correct and incorrect answers, since the score decreases for both correct and incorrect answers. Gaining “points” in this way makes it hard for the user to distinguish between good behavior (correct answers with high confidence) and bad behavior (wrong answers with high confidence), and thus harder to learn to appropriately shift confidence up or down in response to feedback. Previous research has found that scoring rules that reward correct answers by giving positive points and that punish incorrect answers with negative points are more effective at improving calibration than purely positive or purely negative rules (González-Vallejo & Bonham, 2007). Additionally, because the logarithmic score has an unbounded

minimum, a single poor prediction can dominate the total score of a participant, which may be demotivating to a user (Greenberg, 2020). For example, users who receive a score of negative infinity for a single wrong answer with 100% confidence have no incentive to participate further as they can never overcome the effects of their one error.

The Practical scoring rule was designed to fulfill the following characteristics that are hypothesized to create an intuitive and user - friendly scoring mechanism that properly incentivizes forecasters to give accurate judgments (Greenberg, 2020):

1. A higher score should correspond to making a better prediction.
2. There should be a finite upper bound on the number of points a forecaster can get on a single prediction and this maximum should occur when the forecaster makes the best possible prediction (i.e., they are correct and as confident as possible).
3. There should be a finite lower bound on the number of points a forecaster can get in a single prediction (or lose, in which case it is the most negative points possible per prediction). This minimum should occur when the forecaster makes the worst possible prediction (i.e., they are incorrect and as confident as possible).
4. Positive points should only be given if a forecaster makes a correct prediction, and negative points should only be given if they make an incorrect prediction.
5. Forecasters should gain more points for a more confident correct prediction and lose more points for a more confident incorrect prediction.
6. Zero points should be given for predictions with maximum uncertainty (e.g., 50% confidence for a binary choice) since the forecaster provides no information beyond an uninformed prior based on the principle of indifference and thus makes no positive or negative contribution to discrimination.
7. An arbitrarily small change in the prediction should create only a small change in the score (i.e., the scoring function should be continuous with respect to the predicted probability or prediction interval).

Given that the Practical scoring rule is a transformation of the logarithmic scoring rule, it is also a proper scoring rule, so maximizing the Practical score will also maximize calibration. See the below equation,

$$S(p, c) = \frac{s_{max}}{\log(p_{max}) - \log(p_{rand})} * \begin{cases} \log(p) - \log(p_{rand}), & \text{if } c = 1 \\ \log(1 - p) - \log(1 - p_{rand}), & \text{if } c = 0, \end{cases} \quad (1)$$

where p is the forecasted probability, c is whether the forecast was correct or incorrect (correct = 1), s_{max} is the maximum score that a participant can win per forecast (a parameter set to 10 by default), p_{max} is the maximum probability the forecaster is allowed to give as an answer (set to 0.99 by default), and p_{rand} is the probability equivalent to a random guess (e.g., 0.5 for a binary question). The default maximum score of 10 points per question is posited to be an intuitive score for participants to understand and setting the maximum probability to 0.99 avoids the extreme (infinite) negative outliers of a log - based scoring rule.

Note that the Practical scoring shares some of the potential downsides of the logarithmic scoring rule in that it punishes overconfidence more than underconfidence and it can still produce a tail of negative outliers, although to a much lesser degree than the logarithmic rule. The Practical score ranges from -57.27 to 10 on a single trial, which should be comprehensible and not too demotivating for a participant if they get the minimum score. The Practical scoring rule is also not conducive to becoming well - calibrated on extremely common or extremely rare events that occur, for instance, less than 1 in 100 times, since the maximum confidence is capped at 99% (Greenberg, 2020). These downsides are unlikely to be too problematic, given the rule's positive qualities described above. Moreover, a different modified version of the logarithmic scoring rule was shown to have a positive effect on training calibration (Fischer, 1982).

The Practical scoring rule was implemented in an online calibration training app funded by Open Philanthropy (OP) and released on their website in 2018 (Muehlhauser, 2018). The OP app's task consists of judging the relative population of two of the world's largest cities and giving a confidence rating between 50% and 99%. While not explicitly mentioned, this task design is likely inspired by Gigerenzer et al. (1991). These authors argue that confidence judgments about representative samples of a reference class (e.g., city populations from a set of cities) are more ecologically valid than traditional “general knowledge” questions, which are often selected in part for their counter - intuitiveness.

In the OP app, participants are given trial - by - trial feedback with the Practical scoring rule. They can also optionally view their percent correct, cumulative Practical score, and calibration curves after sufficient trials. The app's effectiveness had not been empirically studied during the design phase of our

experiment (S. Greenberg, personal communication, July 19, 2021), so we sought to test a calibration training paradigm inspired by it. We note that Gruetzmacher et al. (2023) independently tested their own modified version of the OP app, which found positive transfer effects to forecasting the outcomes of football games. The differences in the methodologies of our two studies point to important variations in the generalizability of their findings, which we later discuss.

We also note a distinction between calibration for general knowledge questions versus forecasts. Uncertainty for general knowledge questions results from internal uncertainty about facts that can be known with certainty, whereas forecasts involve uncertainty of external events that cannot be known with certainty. Forecasting training involves learning how to map subjective confidence to an appropriate numerical value and learning the predictability of the environment as well as incorporating temporality, among other differences (Wright, 1982; Wright & Ayton, 1986). Benson and Önköl (1992) note that training calibration for forecasting likely requires more training than for general knowledge questions for this reason. It is therefore notable that Gruetzmacher et al. (2023) found positive transfer effects from a general knowledge task to a forecasting task, which has both practical and theoretical implications and motivates further study. For the current study, if a training regime incorporating the Practical scoring rule can be shown to be an effective means of improving calibration on general knowledge questions, it would be suggestive as a tool that could be adapted to training calibration in other contexts, possibly including forecasting.

Our study seeks to test the effectiveness of a performance feedback training paradigm inspired by the OP app that uses the Practical scoring rule as both trial - by - trial feedback as well as block - level feedback. The performance feedback regime also includes the Bias score (Yates, 1990) as block - level feedback. The Bias score is the average confidence minus the percent correct of a set of judgments. Here, it serves as an alternative means of informing participants of their overall calibration, as opposed to the calibration curves used in the OP app, which require large amounts of data to be accurate.

The performance feedback training paradigm of the current experiment is designed to be a relatively short, practical, and testable regime similar to the OP app. It incorporates multiple elements that are believed to be conducive to learning calibration, including both a legible proper scoring rule (the Practical score) and a means of interpreting this score in terms of over - or underconfidence (the Bias score). The performance feedback regime is also similar in spirit to commercially available training

regimes such as the Calibrated Probability Assessments course offered by Hubbard Decision Research (Hubbard Decision Research, 2019), which incorporates multiple methods for improving calibration, including multiple blocks of feedback using the Bias score (Kelly & Mandel, 2024). While the inclusion of multiple training elements in the current design confounds the causal mechanisms behind the results, the design is meant to test a practical tool that includes different elements that plausibly aid in training calibration, rather than test purely theoretical questions related to judgment and cognition. If positive results are obtained, follow - up studies could be done to examine the underlying causal mechanisms and probe more theoretical questions behind the source of the results.

This experiment also compares the effectiveness of the performance feedback paradigm to a simpler outcome feedback paradigm. Our primary hypothesis is that both training regimes will result in better calibration than a control group and that the performance feedback group will be superior to the outcome feedback group. The predictions about outcome feedback are stated with lower confidence, given previous research finding relatively little positive effect of outcome feedback (Lawrence et al., 2006) and one exceptional study that found outcome feedback to be superior to other kinds of performance feedback (Du et al., 2012). The outcome feedback condition of the current study contains strictly less information than the performance feedback condition, so it serves as a useful comparison for the magnitude of improvement due to the enhanced information available in the performance feedback condition. The outcome feedback condition is additionally useful as a contrast with the performance feedback condition since it includes consistent feedback, which might have a positive effect on engagement and thus performance, in contrast to the control condition with no feedback.

Our study also seeks to examine whether individual differences in Actively Open - Minded Thinking (AOMT) predict good calibration and changes in calibration over time. AOMT is a rational thinking disposition that encompasses reflective thinking, tolerance for ambiguity, and seeking information inconsistent with one's prior beliefs (Baron, 1985; Stanovich & Toplak, 2023). Previous studies have shown that AOMT is predictive of measures of rational thinking in many tasks, including good calibration in a forecasting tournament setting (Mellers et al., 2015; Stanovich & Toplak, 2023). Our secondary hypothesis is that AOMT will be positively related to metrics of good calibration and will be predictive of change in calibration over time. That is, if a positive effect of calibration training is found, we can also explore the interactions between AOMT, training condition, and time, with the prediction that higher AOMT would predict increases in performance over time (i.e., a three - way interaction effect). If

successful, this kind of calibration training could be used as a scalable tool to improve judgment accuracy by reducing both overconfidence and underconfidence.

3.7 Experiment 1

3.7.1 Method

3.7.1.1 Participants

We conducted an a priori power analysis using G*Power version 3.1 (Faul et al., 2009) to determine the sample size required to detect a between - subjects effect with a repeated - measures design with six groups (three conditions split by a possible control variable) and six blocks of trials per group. A total sample size of 486 was required to find a small effect size of $f = 0.125$ with power = 0.95, $\alpha = 0.05$ and a correlation between repeated measures = 0.25. The target was increased to 600 participants to account for possible data exclusions and smaller effect sizes.

Participants were recruited from Qualtrics Panels in exchange for Qualtrics credits. Only participants who were above the age of 18, American or Canadian citizens, and spoke English as their first language were allowed to participate in the study. Participants were also required to pre - commit to providing 25 min of undistracted time and to not look up the answers to the general knowledge questions on the internet. Participants were automatically excluded if they gave inconsistent answers to demographic questions about their sex and age at the start and end of the experiment, and if they failed the attention-check question part - way through the experiment (described below). A pilot test of the experiment with 50 participants was conducted to test the methodology and set exclusion criteria before the main set of participants. Post-pilot participants were automatically excluded from the study if they took less than 1/2 the median time of the pilot participants to complete the study (757 s). Finally, post - pilot participants were automatically excluded if they had an average confidence of ≤ 0.55 across all questions, indicating they were likely straight - lining and were not taking the instructions seriously. Given that there were no changes to the methodology besides the speeding and straight-lining exclusion criteria, pilot participants were included in the data analysis, after the post - hoc application of the same exclusion criteria as the post - pilot participants ($n = 6$). A small amount of over - sampling resulted in a total of 610 North American participants (301 Canadian, 305 American, 4 Dual citizens; 306 male sex) being retained after data exclusions.

3.7.1.2 Materials and Procedure

The online experiment was implemented on Qualtrics and participants were disallowed from completing the experiment on a mobile device. Participants first answered demographic questions to determine their eligibility for the study, which included sex (male/female/prefer not to say), age (0–17/18 +), citizenships held (American/Canadian/European/Mexican/ Other), and first language (Cantonese/English/French/Mandarin/Spanish/Other). They were then provided with a brief description of the experiment and gave informed consent. They then committed to providing 25 min of undistracted time and not looking up any answers on the internet.

Participants were randomly assigned in equal numbers to experimental conditions. The first four blocks of the performance and outcome feedback conditions were training blocks. Participants in these conditions were given feedback after each trial and summary feedback after each block (see Stimuli section, below). After the first four blocks, participants in all three conditions then completed an attention check question (described below) before completing two test blocks, where no feedback was provided to any participant. The order of trial stimuli was randomized within blocks and block order was randomized, with the final two blocks always being the test blocks.

Participants were given detailed instructions about what was required of them and the nature of the feedback they would receive before beginning the first trial. In the performance feedback condition, participants were instructed about the nature of the Practical score and Bias score. Participants were told they should seek to maximize their Practical score by keeping their Bias score as close to zero as possible. They were instructed that if they were over - or underconfident in a previous block, they should generally reduce or increase their confidence, respectively, in the next block to get their Bias score closer to zero. In the outcome feedback condition, participants were told only that they would receive feedback about their percent correct answers after each block. In the control condition, participants were informed only about how to complete the experiment.

On each trial, participants made a two - alternative forced choice judgment of the relative population of two urban agglomerations (UAs) drawn from the 100 most populous UAs in 2020 (United Nations, 2019). These stimuli were inspired by the OP app which uses the 2016 population of random world cities, which in turn are likely inspired by Gigerenzer et al. (1991). Participants were randomly assigned to view one of two lists of paired UAs. The UA pairs were generated by creating duplicate lists of the 100 UAs and

randomly sampling without replacement from each list to generate the first and second UA to compare, with the constraint that the two UAs could not be the same. A participant would see one of the 100 UAs between zero and two times over the course of the experiment, with no comparisons between the same two UAs ever occurring.

Between the training and test blocks, participants completed an attention check question entitled “Sports Activities You Engage in Regularly” (Oppenheimer et al., 2009). The attention check contained a long preamble that asked participants not to select any sports from a list presented below the instructions, but to only click on the “next” button instead. Following the preamble, a separate instruction asked participants “Which of these activities do you engage in regularly? (click on all that apply),” followed by a list of popular sports activities. Participants who selected one or more of the sports options were excluded from the experiment.

After the test blocks, participants indicated their level of knowledge for each UA they had seen on a scale from (0) *I have never heard of it before this study* to (4) *I am very knowledgeable about the city*. Then participants completed an eight - item version of the AOMT scale (Cronbach's $\alpha = 0.67$; Baron et al., 2015; Janssen et al., 2020). They rated their agreement with statements related to open - mindedness and critical thinking such as “People should revise their beliefs in response to new information or evidence” on a five - point scale from -2 (strongly disagree) to 2 (strongly agree). Answers were coded such that higher scores indicated more open - minded thinking. Finally, participants were asked about their highest level of education completed (from High School (or equivalent) to University-PhD) and asked again about their sex and their age in years as a means of screening out participants who gave inconsistent answers (Belliveau & Yakovenko, 2022).

3.7.1.2.1 Stimuli

On each trial in all conditions of the main portion of the experiment, participants were presented with a comparison of two UAs in the form “[UA X] had a larger population than [UA Y] in 2020.” Participants were asked to answer True or False and provide a confidence level using a slider from 50% to 100%. After answering a question during training blocks of the experiment, participants in the performance feedback condition were shown whether they were correct or incorrect, a reminder of the confidence of their answer, and their Practical score (see Appendix A for stimulus examples). Participants in the outcome

feedback condition were shown only whether they were correct or incorrect after each question in the training blocks. Participants in the control condition received no trial feedback throughout the experiment.

At the end of each training block of 10 trials, participants in both the performance and outcome feedback conditions were shown a summary table of their performance for all completed blocks. For performance feedback participants, the summary table included their percent correct, the sum of their Practical score, their Bias score, and a verbal interpretation of their Bias score. Given that one of the aims of this research was to design a time - efficient training program, the Bias score was used as the means to inform participants if, on average, they were overconfident or underconfident, since it requires fewer data points to compute a reliable measure of bias compared to the calibration index (Murphy, 1973) or calibration curves (as in e.g., Lichtenstein & Fischhoff (1980)). The verbal interpretation of the Bias score was “over - confident” if $\text{Bias} > 0.01$, “underconfident” if < -0.01 , and “well - calibrated” otherwise. These bounds were chosen based on Lichtenstein and Fischhoff 's (1980) observation that a perfectly calibrated individual will generally produce calibration scores of 0.007 or less, corresponding roughly to a mean Bias of 0.005. Participants in the outcome feedback condition were shown only their percent correct over each completed block in the summary feedback table. Participants in the control condition were only informed that a block was completed.

3.7.2 Results

The experiment was designed to examine the primary hypothesis that calibration could be improved through two kinds of training: performance feedback and outcome feedback. The primary independent variable of the experiment was training condition, with one group for each type of training, plus a control group. The key dependent variables measuring calibration were the sum of the Practical score per block and the Bias score per block. Because the two dependent variables were highly correlated (average transformed Practical score $\times$ average Bias $r = -0.86$; see below for details of the transformation) and they showed the same pattern of results, we report only the transformed Practical score results here (the Bias results are in the online Supporting Information: <https://osf.io/d4ypn/>). Additionally, while the experiment contained a training/test block distinction, analyzing the test blocks in isolation did not provide any additional insights compared to examining the change of transformed Practical score across the entire experiment (null results were obtained). Therefore, we used all training and test blocks for the analysis reported here. However, the test blocks examined in isolation are included in the Supporting Information.

We conducted a mixed design analysis of covariance (ANCOVA) with block as a within - subjects independent variable, training condition as a between - subjects variable and AOMT as a covariate. If training were effective at improving calibration scores, there should be a main effect of condition, with higher average Practical scores in the training conditions compared to the control. There should also be a main effect of block, as the performance in the training conditions would improve during the training blocks whereas the control group would remain constant. Most importantly, there should be an interaction between block and training condition, as performance should increase across blocks more so in the training conditions compared to the control condition.

The secondary hypothesis under consideration was that AOMT would be a significant predictor of calibration. A significant main effect of AOMT would indicate its overall association with calibration. A follow - up regression analysis was performed to assess the directionality of the association, where AOMT was expected to be positively associated with Practical score. The full factorial ANCOVA model also included an interaction effect between AOMT and block. A significant interaction effect between AOMT and block would indicate that AOMT was beneficial to *learning* calibration if Practical score increased across blocks, averaged across treatment condition. If a significant interaction effect between block and training condition were found in support of the primary hypothesis, follow - up analyses examining higher-order interactions between block, condition, and AOMT could be performed. A significant three - way interaction between block, condition, and AOMT would be indicative of AOMT facilitating learning calibration in the training conditions, if the effect was present in one or both of the training conditions, but not the control group. Where nonsignificant results were obtained, follow - up Bayesian analyses were conducted to assess support for the null hypothesis.

All results were calculated using JASP version 0.17.2.1 (Jasp Team, 2023). Figures were generated using the raincloudplots package in R version 4.0.0 (Allen et al., 2021; R Core Team, 2021). Where the assumptions of the ANCOVAs were violated beyond what could be overcome by the large, evenly distributed sample, either transformations to the data were made or nonparametric tests were conducted and are discussed where relevant. The Greenhouse–Geisser correction was used when Mauchly's test of sphericity was significant at the $p < 0.05$ level.

3.7.2.1 Preliminary Analyses

We conducted analyses to check the assumptions behind the main analyses (see online Supporting Information for details). To summarize, the two UA lists participants were assigned to had no discernible effect on performance, so the two lists were combined. AOMT did not differ between the three training groups, so it was appropriate to include it as a covariate. Participants recognized an average of 63.80% of cities they saw ($SD = 0.18$, 95% CI [0.62, 0.65]) and reported an average knowledge of 1.20 ($SD = 0.61$, 95% CI [1.15, 1.25]) on a scale from 0 to 4, indicating that while largely unfamiliar with the UAs, they were knowledgeable enough to do the task. Mean accuracy across the six blocks was statistically above chance ($M = 0.60$, $SD = 0.08$, 95% CI [0.59, 0.60]). Median completion time was 1499.50 s (IQR = 810.50 s).

3.7.2.2 Practical Score Across Blocks

The Practical score variable across all six blocks showed considerable negative skew (range: -1.63 to -3.34) and leptokurtosis (range: 3.25–19.30). This is likely because the data were generated with a logarithmic scoring rule that can produce a long tail of negative outliers. Removal of the most extreme low - scoring participants was not desirable, since even the most extreme low-scoring participants ($z \leq -3.09$) were above chance accuracy ($M = 0.56$, $SD = 0.05$, 95% CI [0.53, 0.58]). Their removal may thus have implied removing compliant participants.

To address the issue of negative skew and leptokurtosis, the data were exponentially transformed with base 100. An exponential transformation was appropriate given that the data were negatively skewed and followed a reversed lognormal distribution (Feng et al., 2014). However, an exponential transformation of the raw scores yielded an extreme range of numbers that spanned more than 200 orders of magnitude. To make the computations tractable, the data were first min - max scaled before being exponentially transformed. The base 100 was chosen for its trade - off in minimizing skew and kurtosis and maximizing interpretability, as the transformed scores fell on a scale between 1 and 100 (where 100 was also the maximum non - transformed score within a block). The transformed data showed much less skew (range: -0.63 to 0.52) and leptokurtosis (range: -0.01 to 3.61) than the original Practical scores.

The Greenhouse–Geisser corrected ($\epsilon = 0.98$) ANCOVA with the design explained above revealed no statistically significant main effect of training condition, nor was there a main effect of block, nor were there any interaction effects, all $p > 0.10$. This is contrary to the primary hypothesis that there would be main effects of condition and block as well as an interaction effect between these two variables (see Figure

3.1 and Supporting Information Table S1 for full ANCOVA results). In partial support of the secondary hypothesis that AOMT would be predictive of Practical score and interact with block, there was a statistically significant main effect of the AOMT covariate, $F(1, 606) = 37.50, p < 0.001, \eta_p^2 = 0.06$. As was predicted, a follow - up regression analysis showed that AOMT was positively related to the average transformed Practical score, $b = 0.45, t = 6.20, p < 0.001, 95\% \text{ CI } [0.31, 0.60]$.

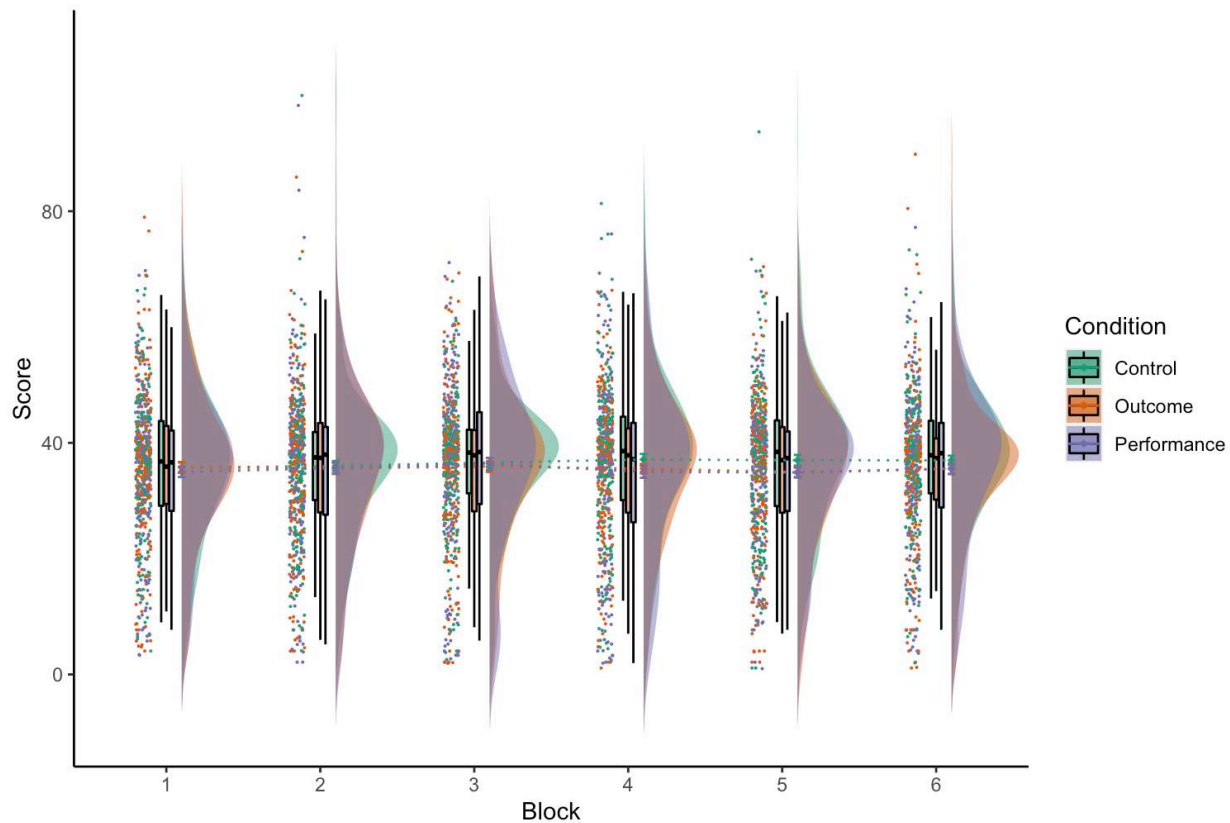


Figure 3.1. Min-max scaled exponentially transformed practical score by block and condition in Experiment 1.

Note. Error bars around the mean represent one standard error of the mean.

We also conducted a Kruskal–Wallis test of the effect of training condition on the average untransformed Practical score of all six blocks, which found no statistically significant effect of training condition, $H(2, N = 610) = 1.98, p = 0.37$. In sum, the results did not support the primary hypothesis that calibration training would lead to an increase in Practical score over the course of the experiment, relative to the control condition. The secondary hypothesis that AOMT would be positively related to Practical score overall was supported, but it did not predict change in Practical score across blocks. Of course, the latter result is

unsurprising given that there was no significant change across blocks. Given the lack of support for the primary hypothesis, we did not investigate higher - order interaction effects with AOMT.

To explore possible explanations for the null results found above, we analyzed the subcomponents of the Practical score (accuracy and confidence; full results are available in the online Supporting Information). Mean accuracy did not differ significantly across groups, nor did it change significantly across blocks, nor was there an interaction between the two (all $ps > 0.10$). Confidence also showed no significant differences between groups, no significant change across blocks, and no interaction between the two (all $ps > 0.10$). These results provide evidence that the training had no effect, as opposed to a more complex interaction between accuracy and confidence, for example.

3.7.2.3 Bayesian Analyses

Given the presence of null results, follow - up Bayesian statistical analyses were conducted with JASP to quantify the degree of evidence in favor of the null versus alternative hypotheses using Bayes factors (BFs). The default JASP settings for model specifications were used, except the numerical accuracy and posterior samples variables, which were set manually to 10,000 and 1000, respectively, to reduce computation time. Details about the JASP analyses and results can be found in the online Supporting Information.

We report the BF_{01} s of a Bayesian ANOVA model equivalent to the frequentist one reported above, with training condition and block order as the independent variables (the AOMT covariate was omitted for simplicity of interpretation; see Table 3.1). The BF_{01} s represent the extent to which the data favor the null model, where the difference in means is assumed to be zero, compared to an alternative model with some combination of condition, block, and their interaction as predictors. A $BF_{01} > 1$ indicates the evidence favors the null model over the alternative model, where larger numbers indicate more evidence in favor of the null model. Following van Doorn et al. (2021), BF_{01} s between 1 and 3 are considered weak/inconclusive evidence in favor of the null model, between 3 and 10 are considered moderate evidence, and greater than 10 are considered strong evidence. We also report the error percentage of the BF_{01} , which represents the uncertainty of the BF_{01} due to the Markov chain Monte Carlo (MCMC) procedure used to generate it. A BF_{01} of 10 with a 20% error percentage would indicate that the BF_{01} is expected to fluctuate between 8 and 12. Error scores below 20% are considered acceptable (van Doorn et al., 2021).

Table 3.1. Experiment 1 Bayesian ANOVA results.

Models	BF ₀₁	Error %
Condition	28.38	0.84
Block	17141.39	1.66
Block + Condition	492883.79	1.19
Block + Condition + Block x Condition	6.77×10 ⁺¹⁰	1.33

Note. All models include subject and random slopes for all repeated measures factors.

The results suggest strong support for the null hypothesis that there was no main effect of condition, block order, nor an interaction between them. This suggests that the null results found above were not due to a lack of power to detect true differences, but rather due to a lack of true differences. The small effect sizes of the frequentist analyses would further indicate that there was no substantial effect of training present.

3.7.3 Discussion

Experiment 1 failed to support the primary hypothesis that either training condition would increase the Practical score compared to the control group. The secondary hypothesis that AOMT would be positively related to Practical score was supported, although AOMT did not interact with block, likely because Practical score did not significantly vary across blocks.

One possible reason for the results is that 10 trials per block were not enough to produce a reliable measure of Bias and Practical score for the block - level feedback. For instance, a typical participant's hit rate of ~60% would yield a standard deviation of 1.55 hits per block of 10 trials, meaning the Bias score of a well - calibrated participant would be between ± 0.155 , on average. If participants were receiving highly variable feedback about their performance in the previous block, it may have led to adjustments that added more noise than signal to their performance. It may have also inadvertently produced an illusion of control, where participants thought they were improving their calibration, but the quality of information they were receiving was not high enough to lead to real changes. It is also possible that 40 total training trials were insufficient practice to improve calibration.

Participant engagement was also an issue in this experiment, since at least some of the participants were not above chance accuracy in the test blocks (see outlier analysis in the Supporting Information). While

low accuracy is not a problem per se in calibration training, since participants could adjust their confidence to match their low accuracy, it makes it difficult to distinguish noncompliant participants from those who were simply not knowledgeable about world UAs. Experiment 1 used an exclusion criterion for straight - lining of average confidence ≤ 0.55 , which may have been too stringent given that mean accuracy was around 60%. Experiment 2 was designed to address these issues.

3.8 Experiment 2

3.8.1 Method

Experiment 2 followed the same methodology as Experiment 1 except for the changes described here. Forty new UA pairs were generated for the two UA lists according to the methodology in Experiment 1. All 100 UA pairs were reconsolidated into five blocks of 20 trials. There was only one test block of 20 trials in Experiment 2. After completing the test block, participants were asked the following question to gauge their engagement in the study: “Did you find the task engaging enough so that you were able to pay attention for the entire time? (You will not be penalized based on your answer).” Participants who answered “No” to this engagement question were allowed to continue the study but were removed from the data analyses ($n = 39$; their characteristics are explored in the Supporting Information). To make the experiment shorter, participants did not answer familiarity questions after the main task.

We anticipated the removal of up to a third of participants due to their answer to the engagement question and we wanted to increase the power of Experiment 2, so the target sample size was increased to 900. Due to a small amount of over - sampling, a total of 944 participants with the same demographic profile as Experiment 1 were recruited from Qualtrics Panels. As in Experiment 1, a pilot test with 47 participants was conducted before the main set. Due to a bug in the code, some of the pilot participants in the outcome and performance feedback conditions received the wrong feedback for one question. The bug was fixed and pilot participants who were potentially exposed to the bug were removed from the analysis post - hoc ($n = 16$). The remaining pilot participants were included in the analysis, after post - hoc application of the same exclusion criteria applied in the main set of participants. Post - pilot participants were automatically excluded from the experiment if they took less than 700 s to complete the study, which was less than 1/2 of the median completion time of a pilot study with 47 participants. Post-pilot participants were also automatically excluded from the study if their average total confidence was ≤ 0.51 or ≥ 0.99 . However, because the Qualtrics software censored confidence values of 100% to a value slightly below

0.99, all participants with average confidence ≥ 0.989 were removed post-hoc. After all data exclusions, 871 participants were retained.

3.8.2 Results

3.8.2.1 Preliminary Analyses

As in Experiment 1, there was no discernible difference between the two UA lists, and AOMT did not differ between training groups, so the UA lists were combined and AOMT was used as a covariate. Average accuracy across the five blocks was statistically above chance ($M = 0.59$, $SD = 0.07$, 95% CI [0.59, 0.60]). Median completion time was 1608 s (IQR = 802 s).

3.8.2.2 Practical Score Across Blocks

For the same reasons as in Experiment 1, we transformed the raw Practical score values in each block using the exponential function with base 200, after the scores had first been min - max scaled. The resulting transformed values were much less negatively skewed (skew range: -0.90 to -0.32), and less leptokurtotic (kurtosis range: 0.19 to 1.97) than the original scores.

We conducted a Greenhouse - Geisser corrected ($\epsilon = 0.98$) ANCOVA on transformed Practical score with the same design as Experiment 1. Contrary to the primary hypothesis that there would be main effects of condition and block as well as an interaction effect between these two variables, the results revealed no significant main effect of training condition, nor was there a main effect of block order, nor were there any interaction effects, all $p > 0.10$ (see Figure 3.2 and Supporting Information Table S2 for full ANCOVA results). In partial support of the secondary hypothesis that there would be a main effect of AOMT and it would interact with block, there was a statistically significant main effect of AOMT score, $F(1, 867) = 18.54$, $p < 0.001$, $\eta_p^2 = 0.02$. A follow - up regression analysis found that AOMT was positively linearly related to the average transformed Practical score, $b = 0.54$, $t = 4.27$, $p < 0.001$, 95% CI [0.29, 0.79], as was predicted.

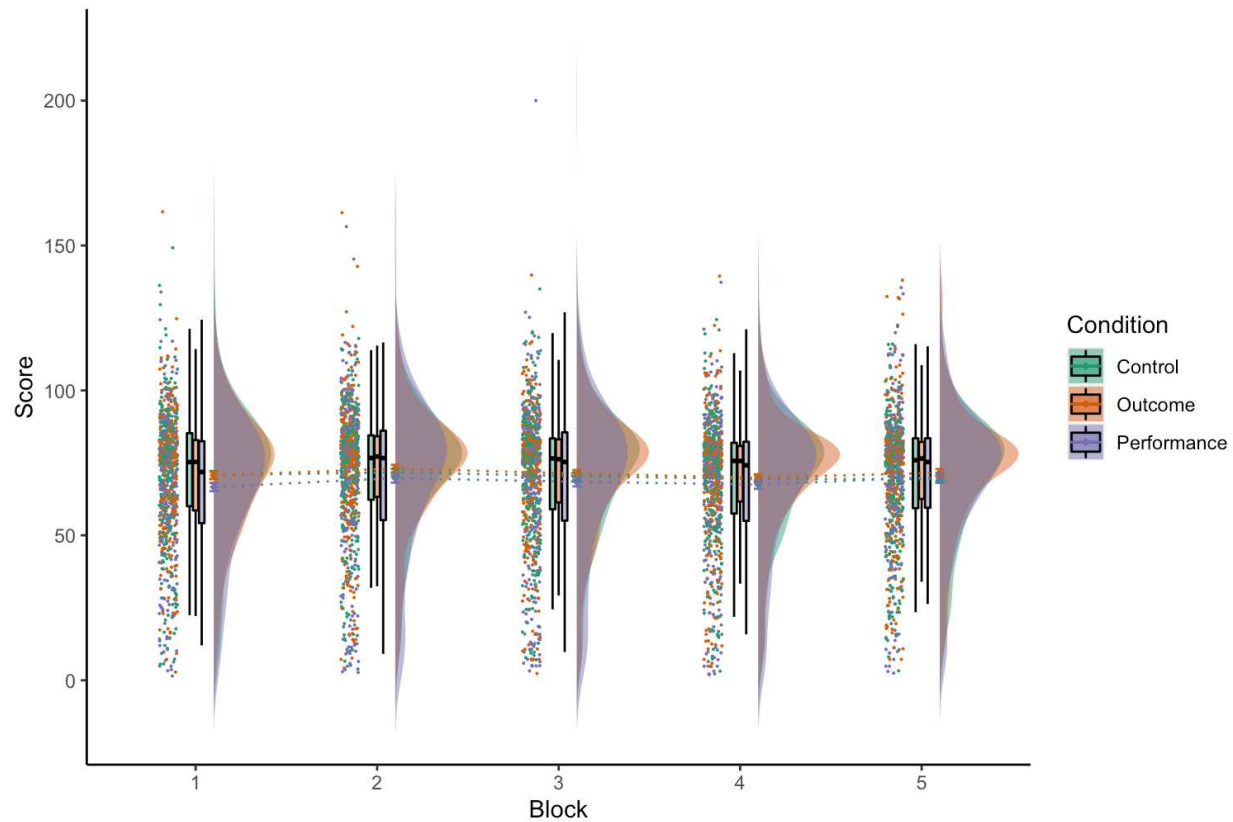


Figure 3.2. Min-max scaled exponentially transformed practical score by block and condition in Experiment 2.

Note. Error bars around the mean represent one standard error of the mean.

We also conducted a Kruskal–Wallis test of the effect of training condition on the average untransformed Practical score of all five blocks, which found no statistically significant effect of condition, $H(2, N = 871) = 4.62, p = 0.10$, contrary to the primary hypothesis. Thus, the results did not support the primary hypothesis that calibration training would lead to an increasing Practical score relative to the control condition. The secondary hypothesis that AOMT would be positively related to overall Practical score was supported, but AOMT did not predict the change in Practical score across blocks. Due to the overall lack of effect of training condition, we did not explore higher - order interactions between training condition, block, and AOMT.

To explore possible causes of the null effects, we conducted exploratory analyses on accuracy and confidence. For mean accuracy, there was no significant main effect of condition, block, nor an interaction between the two (all $ps > 0.10$). For mean confidence, there was a significant main effect of condition ($p =$

0.009), a significant main effect of block ($p < 0.001$), but no significant interaction between them ($p = 0.346$). Descriptively, average confidence was highest in the performance feedback condition and all groups' confidence decreased over the course of the experiment. These results do not support the claim that training had a positive effect on calibration (due to the lack of interaction effect) but rather provide some evidence that the performance feedback training may have had a deleterious effect, as it had the highest overall confidence, contrary to the prediction that the performance feedback condition would show the lowest (over)confidence. This statement is caveated by the fact that raw confidence was not a target variable for the intervention and that the effect was not seen in the target variables of Practical score or Bias (full results for each of the four variables discussed are available in the online Supporting Information, as are the results of test blocks analyzed in isolation).

3.8.2.3 Bayesian Analyses

The results of the transformed Practical score were also analyzed using Bayesian methods following the same procedure described in Experiment 1 (see Table 3.2). The results strongly favored the null hypothesis that there was no main effect of training condition, no main effect of block order, nor an interaction between them.

Table 3.2. Experiment 2 Bayesian ANOVA Results

Models	BF ₀₁	Error %
Condition	11.20	0.51
Block	55.08	0.91
Block + Condition	621.38	0.82
Block + Condition + Block x Condition	1.40×10 ⁺⁷	4.00

Note. All models include subject and random slopes for all repeated measures factors.

3.8.3 Discussion

Experiment 2 failed to support the primary hypothesis that calibration training would lead to higher Practical scores compared to the control group. The secondary hypothesis that AOMT score would be positively related to Practical score was supported, although AOMT did not predict change in Practical score over blocks. This is likely because there was no significant change in Practical score over blocks.

3.9 General Discussion

Both experiments of our study failed to support the primary hypothesis that training using either outcome feedback or performance feedback would have a positive effect on the calibration of confidence. We were surprised by this result, given previous research showing positive effects of performance feedback training (Benson & Önköl, 1992; Lichtenstein & Fischhoff, 1980; Stone & Opel, 2000), especially that of scoring rules that reward positive and punish negative behavior (González-Vallejo & Bonham, 2007). Indeed, our results contradict the only other known study to investigate the Practical scoring rule as a form of performance feedback (Gruetzmacher et al., 2023), as discussed further below. Despite the failure to support the primary hypothesis that calibration training would be effective, the results did partially support the secondary hypothesis that AOMT would be positively related to Practical score, as previous results would suggest (Mellers et al., 2015). AOMT did not predict change in Practical score over blocks, but this is unsurprising given the lack of change in Practical score in general.

While the lack of effect of our performance feedback regime was unexpected, it is worth noting that the literature shows that successful calibration training regimes are difficult to achieve, with many studies finding either mixed or negative results (e.g., Du et al., 2012; Sieck & Arkes, 2005; Stone et al., 2023). For instance, a commercial calibration training course that combined calibration feedback and exposure to various debiasing strategies appeared to reduce confidence, but not in a manner that was sensitive to whether overconfidence or underconfidence was exhibited (Kelly & Mandel, 2024). It is perhaps less surprising that outcome feedback was ineffective at improving calibration, as studies tend to show this is the least effective form of feedback (Lawrence et al., 2006; Önköl & Muradoğlu, 1995), although there are exceptions to this general rule (e.g., Du et al., 2012; Subbotin, 1996).

To our knowledge, Gruetzmacher et al. (2023) is the only study to examine the effectiveness of a calibration training paradigm based on the Practical scoring rule. In their experiment, participants completed either 60 or 120 training trials of comparing the populations of American cities with the Practical scoring rule as feedback, followed by a forecasting tournament about the outcomes of local football games. Unlike our study, no other calibration training or feedback was provided. They found that in the 120 - question group, participants were nearly perfectly calibrated in forecasts between 50% and 70% confidence and this group was overall better calibrated than the control group. They attribute the superior calibration of the 120 - question group to the fact that they received a set of 10 hard questions after the initial 60, followed by questions of diminishing difficulty, and not to the longer training. Our

findings challenge this explanation since our study contained difficult questions that elicited confidence in the 50%–70% range but found no effect of training. Given the trajectory of the results of our experiments, it is also doubtful that longer training explains the difference in findings. The greater variance in difficulty and some of the other differences in the stimuli between the studies may explain the discrepancy, as do differences in the analysis of their results.

Beginning with the discrepancies in how they analyzed their data, Gruetzmacher et al. (2023) used a coarsened exact matching technique (Iacus et al., 2012) that balanced the control and treatment groups on the variables of the Cognitive Reflection Test (CRT) score (Frederick, 2005) and self - reported expertise in college football. They also included gender as a factor in their model, whereas we did not. It is unclear to what extent these factors played a role in the discrepancy in the findings of the two studies, although we question their use of a matching technique intended for observational data in a study where random assignment was used. We explored the effect of including sex as a factor in our model, which did not substantially change the results (this exploratory analysis is available in the Supporting Information). We did not measure CRT or self - rated expertise, so we cannot balance our treatment groups using these variables to examine their effects on our results.

Regarding the differences in stimuli, our task was notably harder than the task used by Gruetzmacher et al. (2023). The latter used American cities primarily located in the region where participants lived, as opposed to global UAs, as in our study. Their city populations ranged across three orders of magnitude, whereas our UAs only spanned two, which would also make our stimuli more difficult to distinguish. The higher familiarity of their stimuli likely provided a wider and more homogeneous set of valid contextual cues to draw upon, leading to overall better accuracy and calibration. Furthermore, the higher difficulty of our task combined with extremeness aversion (where participants tend to avoid the extreme ends of a scale; Simonson & Tversky, 1992; Teigen, 1983) may have produced a ceiling effect in which participants were unwilling to select values near 50% despite being completely uncertain.

Gruetzmacher et al. (2023) also increased participants' compensation based on performance, which may have increased their participants' motivation relative to ours. Our study was completed entirely online, which arguably created an incentive for speed rather than diligence. Finally, Gruetzmacher et al. did not include a pedagogical element relating to what calibration is or how to achieve it, nor did they provide information about Bias or global over - or underconfidence, as in our study. It is possible that the inclusion

of the Bias metric and its verbal interpretation at the end of each block of our study was counterproductive and undid the positive effects of the Practical score. We argue against these possible explanations in turn.

First, while our task was difficult, participant accuracy was near 60% and statistically above chance. Other studies, most notably Lichtenstein and Fischhoff (1980) and González-Vallejo and Bonham (2007; in their zero - repetition condition most similar to ours) have found positive effects of performance feedback despite accuracy near 60% (or less) while also using a 50%–100% scale, so this seems unlikely to be a substantial cause of the lack of effect. These two studies, along with others in the relevant literature (Koriat et al., 1980; Zakay, 1992), used general knowledge stimuli that were not derived from a representative sample of a reference class. Accordingly, the potentially greater variety of valid cues available to Gruetzmacher et al.'s (2023) participants compared to ours is also unlikely to be a main cause of our null findings. The greater variance of difficulty in Gruetzmacher et al.'s stimuli may have had an effect and warrants further study.

Second, while our online participants were arguably incentivized to complete the task quickly instead of carefully, researchers have found that online samples replicate many psychological findings found using conventional sources and show similar psychometric properties (Behrend et al., 2011; Buhrmester et al., 2011; Feitosa et al., 2015; Paolacci et al., 2010). For example, a recent meta - analysis of over 90 studies in psychology research using commercial online panel data found that the results of online samples have similar psychometric properties and fall within meta - analytic confidence intervals of previous studies using conventional data, so long as methodological precautions against low - quality data are taken (Walter et al., 2019). Arguably online samples are more representative of the wider population than convenience samples of college students (Paolacci et al., 2010; Walter et al., 2019). Online samples tend to be older, more diverse, and have more work experience than undergraduate convenience samples (Behrend et al., 2011), although they tend to be younger, more diverse, higher educated, but more poorly paid than the general US population (Paolacci et al., 2010). Our study followed best practices for online participant data screening, including using only native English speakers, excluding participants based on speeding, providing inconsistent responses, failing attention checks, or straight - lining/giving nonsensical answers, in addition to the default Qualtrics data integrity procedures that protect against bots, multiple submissions from one participant and other fraudulent behaviors' (Belliveau & Yakovenko, 2022). Additionally, our study replicated the finding that AOMT is predictive of good calibration (Mellers et al., 2015), which adds to our confidence that our participants were not merely providing noise.

While we are reasonably confident that our participants were engaged in the task enough to pass the screening criteria and were not merely providing noise, we did not probe participants' understanding of the instructions or the informativeness of the feedback they received. Examining this question would be a worthwhile aspect to explore in future research, especially given that many of the relevant studies in the domain do not probe participants' understanding of the instructions or feedback (e.g., González-Vallejo & Bonham, 2007; Gruetzmacher et al., 2023; Stone et al., 2023). This is noteworthy given that the amount of pedagogical training varies widely between studies, ranging from no pedagogical instructions at all (Gruetzmacher et al., 2023) to automated personalized feedback (Stone et al., 2023), to intensive in-person feedback sessions (Lichtenstein & Fischhoff, 1980). Examining how the type of pedagogical feedback impacts participant comprehension or motivation and how that in turn affects the effectiveness of the training would be valuable information in future studies.

Finally, regarding the possible detrimental effect of including the Bias score, other studies have included it as a feedback metric in addition to other calibration metrics in successful training experiments (Lichtenstein & Fischhoff, 1980; Önköl & Muradoğlu, 1995). Our participants were told to use the Bias metric and its verbal interpretation to help maximize their Practical score. It is possible that the Bias metric was too noisy as a signal to participants, since it would indicate they were globally over- or underconfident unless they were very precisely well-calibrated. Given that this effect would produce a noisy but unbiased signal, and that participants were consistently overconfident, it is unlikely that the inclusion of the Bias metric had a countervailing negative effect on a positive effect of the Practical score. The most likely hypothesis to explain our results is that calibration training using the Practical score is not as robustly positive as Gruetzmacher et al. (2023) would indicate, and more research is needed to determine the conditions under which it is effective.

3.9.1 Conclusion

In conclusion, we found no positive effect of calibration training using either a performance feedback regime with the Practical scoring rule or outcome feedback. This result is contrary to the only other known study to examine the Practical scoring rule as a form of performance feedback (Gruetzmacher et al., 2023), and contrary to another similar scoring rule paradigm (González-Vallejo & Bonham, 2007). A plausible reason for the lack of positive effect was the high difficulty level of our task, which was based on the original OP app which uses world city populations as forced choice stimuli. Our findings cast some doubt on the robustness of the Practical scoring rule as an effective form of performance feedback training.

Future research should examine the conditions under which such a training paradigm is effective, particularly under what difficulty level and its variability. Additionally, more research should examine the effect of including a pedagogical element and the Bias score along with the Practical score on the effectiveness of such performance feedback training. Finally, once more robust techniques for training calibration under this kind of performance feedback are found, future research should examine how other individual differences influence learning calibration, such as the Rational - Experiential Inventory (Pacini & Epstein, 1999) or CRT score (Frederick, 2005) as this remains an under - explored area of research and could be conducive to determining which individuals are more amenable to training.

The Practical scoring rule satisfies many desiderata of an effective feedback mechanism to train calibration. More research should be conducted to examine the conditions under which it is effective and the mechanisms that underlie its effectiveness. A brief and effective calibration training tool that could be automated would be a valuable resource for governments, businesses, and private actors to make more informed predictions and decisions that could help them avoid costly mistakes or allocate appropriate resources to promising projects (Hernandez, 2017; Mandel & Irwin, 2021). Conversely, options for such training must be carefully evaluated within multiple experimental paradigms to avoid overgeneralization from potentially weakly replicable successes. This is not only important for the scientific record but also so that consumers of such products may be well informed before allocating their finite resources to pursue particular strategies purported to have ameliorative effects on judgment quality.

3.10 Acknowledgments

We wish to thank Daniel Irwin for his assistance with this research. This research was supported by a doctoral training scholarship to the first author by the Fonds de recherche du Québec—Nature et technologies (#303239). It was also supported by the Accelerate Command, Control, and Intelligence Project Activities #AC2I - 019 and AC2I - 031 and Canadian Safety and Security Program project #2018 - TI - 2394 under the direction of the second author. This research contributes to the North Atlantic Treaty Organization Systems Analysis and Studies Panel Research Task Group on Anticipatory Intelligence for Superior Decision - Making (SAS - 189) under the co - direction of the second author.

3.11 Conflicts of Interest

The authors declare no conflicts of interest.

3.12 Data Availability Statement

The data that support the findings of this study are openly available in the Open Science Framework repository at <https://osf.io/d4ypn/>.

3.13 Supporting Information

Additional supporting information can be found online in the Supporting Information section.

CHAPTER IV

Improving Geopolitical Forecasts with Bayesian Networks

4.1 Overview of Chapter 4

The experiments in Chapter 3 were designed to find a short-duration training paradigm to improve calibration of subjective confidence. If successful, this kind of paradigm could be incorporated into forecasting training for providing better-calibrated probabilistic forecasts. Additionally, not all probabilistic judgments relevant to the intelligence community (IC) are predictions. Explanatory judgements about past or present facts, situations, or developments that the analyst could know in principle but lacks sufficient information for certainty would also benefit from well-calibrated probability estimates (Barnes, 2016).

Training and practice are not the only way to improve accuracy of subjective probability judgements. As the Good Judgement Project (GJP) found, algorithmic recalibration and aggregation of forecasts could improve accuracy over and above training and practice effects (Tetlock et al., 2014). Correcting judgements post hoc may prove to be a superior means of improving accuracy than through SATs or training, as algorithms may be better at eliminating bias than human reasoners (Kuncel et al., 2013). However, the positive benefits of using analytic techniques valid to the task at hand, trained intuitions, and algorithmic improvements likely compound, as the GJP's investigations into the effects of probability training, practice, and algorithmic recalibration and aggregation would suggest (Chang et al., 2016; Ungar et al., 2012).

Multiple forms of algorithmic improvements to forecasts have been shown to be effective, ranging from recalibration and aggregation (Atanasov et al., 2017), to statistical methods like regression (Cross et al., 2018), to machine learning (Shinitzky et al., 2024). Chapter 4 uses Bayesian networks (BNs) and logistic regression in conjunction with recalibration and weighted aggregation to explore if these machine learning techniques can improve the predictive accuracy of recalibration and aggregation methods alone. (Atanasov et al., 2017; Shinitzky et al., 2024). If successful, this methodology could improve on the state of the art of algorithmically enhancing the accuracy of probabilistic estimates, which could be applied in forecasting and other situations where probabilistic estimates are available.

4.2 Improving Geopolitical Forecasts with Bayesian Networks

Matthew Martin

Université du Québec à Montréal

Montreal, Quebec, Canada

Correspondence: martin.matthew@courrier.uqam.ca

201, avenue du Président-Kennedy

Local PK 4150

H2X 3Y7 Montréal (Québec)

Publication status: Submitted to the International Journal of Forecasting on January 6, 2026.

4.3 Abstract

This study explores how Bayesian networks (BNs) can improve forecast accuracy compared to logistic regression and recalibration and aggregation methods, using data from the Good Judgment Project. Regularized logistic regression models and a baseline recalibrated aggregate were compared to two types of BNs: structure-learned BNs with arcs between predictors, and naive BNs. Four predictor variables were examined: absolute difference from the aggregate, forecast value, days prior to question close, and mean standardized Brier score. Results indicated the recalibrated aggregate achieved the highest accuracy (AUC = 0.985), followed by both types of BNs, then the logistic regression models. Performance of the BNs was likely harmed by reduced information from the discretization process and violation of the assumption of linearity likely harmed the logistic regression models. Future research should explore hybrid approaches combining BNs with logistic regression, examine additional predictor variables, and account for hierarchical data dependencies.

Keywords: Bayesian Networks, Machine Learning, Forecasting, Logistic Regression, Recalibration, Aggregation, Judgment

4.4 Introduction

Governments, militaries, businesses, and other stakeholders must make decisions and long-term plans in the face of uncertainty about the future state of the economy, the actions of other states, and world events. Anticipating important geopolitical developments allows stakeholders to make and execute strategic plans to better achieve their goals. Increasing the accuracy of geopolitical forecasts thus has immense value to these actors. To this end, the US government has funded several projects related to improving geopolitical forecasting, including the Aggregative Contingent Estimation (ACE) tournament in 2010, in which teams competed to find the best means of eliciting, aggregating, and representing probabilistic judgements of future geopolitical events (IARPA, 2010). The winning team was the Good Judgement Project (GJP) led by Phil Tetlock and Barbara Mellers, who used participant training, teaming, and forecast recalibration and aggregation techniques to achieve accuracy superior to professional intelligence analysts who had access to classified information (Goldstein et al., 2015; Good Judgment Inc., 2025).

Since the ACE tournament, national governments have begun using similar forecasting techniques on an ongoing basis, such as the UK's Cosmic Bazaar (The Economist, 2021) and the RAND forecasting initiative, which advises the US government (RAND, 2025). Businesses have also begun to use forecasting techniques to predict the outcomes of their internal projects and competitive environments, such as Twitch and Google (Hernandez, 2017; Schwarz, 2024). Organizations such as Metaculus and Good Judgement Open also provide forecasting platforms that allow questions about a diverse range of topics for a wide range of clients (Good Judgment Open, 2025; Metaculus, 2025).

Researchers have discovered several psychological variables predictive of forecasting accuracy. According to Mellers et al. (2015), the most important are measures related to a forecaster's behaviours such as the number of forecasts per question and the amount of time spent deliberating per question. Situational variables such as training in probabilistic reasoning and being a member of a team are also important factors predicting accuracy, as are dispositional variables such as a general intelligence, political knowledge, and open-minded thinking (Mellers et al., 2015).

Atanasov and Himmelstein (2023) conducted a separate review of the predictors of forecasting accuracy. In summary, the authors suggest that the metrics most associated with an individual's success in predicting outcomes are, in order of descending value: accuracy-related (e.g., accuracy on previously resolved questions), intersubjective (e.g., squared distance from a consensus forecast), behavioural (e.g., frequency

of updating), dispositional (e.g., intelligence and personality), and then finally expertise-related (e.g., knowledge tests). The authors suggest that these metrics would serve as useful means of giving higher weight to forecasters that exhibit these characteristics.

Relatedly, researchers have explored various ways of using algorithmic and statistical techniques to enhance the information available from raw forecasted probabilities to predict geopolitical events. These techniques range from statistical techniques like regression (Cross et al., 2018), to aggregation and recalibration (Atanasov et al., 2017), and machine learning (Shinitzky et al., 2024). To this end, Shinitzky et al. (2024) explored the extent to which standard machine learning techniques can improve the accuracy of geopolitical forecasts. They tested regularized versions of Bernoulli naive Bayes, gradient boosting classifier, linear discriminant analysis, and logistic regression on their ability to predict binary event outcomes given participants' forecast probabilities and over 40 other predictor variables.

The additional predictor variables included variables derived from the group's forecasted probabilities (e.g., the absolute difference of a forecast from the mean), other features elicited from the forecasters (such as which dimensions they considered in making their prediction), linguistic features from the forecast rationales, features about the participants' psychology (e.g., locus of control), and demographics (e.g., education level), as well as ratings provided by external evaluators about the forecasts' convincingness. Shinitzky et al. (2024) found that regularized logistic regression had the highest performance of all machine learning techniques they tested, as measured by area under the receiver operating characteristic curve (AUC). Using regularized logistic regression, they then tested combinations of their features to see which were most predictive of the outcomes of geopolitical events. They found that the extremeness of the forecasted probability and absolute distance from the group consensus were the most important variables, followed by linguistic features of the forecast rationales. Finally, they used the classification probabilities from the winning logistic regression model as weights in an aggregation procedure, showing that the probability weights increased the accuracy of an aggregate prediction, compared to a simple confidence-weighted baseline aggregate.

While Shinitzky et al. (2024) helpfully explored how standard machine learning techniques can be applied to improve forecasting accuracy, questions can be raised about how replicable their results would be in contexts more commonly found on forecasting platforms such as Good Judgement Open or Metaculus. The first consideration is that their sample was relatively small, consisting of only 153 participants and 10

questions. For comparison, the original ACE GJP tournament consisted of over 9000 participants and 600 questions. Part of the rationale behind prediction platforms is to leverage the wisdom of the crowd, and thus maximizing the number of forecasters, questions, and answers per question is an important goal. A second consideration is that the format of Shinitzky et al.'s (2024) tournament did not use accuracy metrics as predictors in their main analysis, as no questions had resolved at the time of forecast elicitation. While there are certainly situations where such 'cold-start' conditions hold, accuracy-related scores are the best class of predictors for future accuracy and are an important part of forecast aggregation procedures (Atanasov et al., 2017; Atanasov and Himmelstein, 2023). A third consideration is that for 9/10 of the questions, only one forecast per participant was elicited, with only one question allowing for a single update. This is problematic because some of the non-accuracy-related metrics most predictive of accuracy are the frequency and magnitude of updates to forecasts (Atanasov and Himmelstein, 2023). A fourth consideration is that the time range of Shinitzky et al.'s (2024) forecasting period was relatively restricted, where predicted events were between one and four months in the future. In comparison, the GJP tournament had forecasts ranging from 1 to 549 days in the future. In longer term forecasting situations, the change in estimated probability over time is often decision-relevant, as it can signal important geopolitical developments necessitating strategic choices. Finally, some of the metrics used as predictors of outcomes in Shinitzky et al.'s (2024) study would be impractical to implement in large-scale settings, such as third-party evaluations of the convincingness of other forecasters' rationales.

Given the importance of improving geopolitical forecasting, researchers should examine alternative approaches and whether Shinitzky et al.'s (2024) results generalize to forecasting contexts more similar to those used by governments, where there is a wider variety and number of participants, forecasts per participant, and timing of forecasts. These contexts also allow for examining forecaster accuracy and other behaviours such as update frequency. To this end, the current study uses data from the original GJP ACE tournament (Good Judgment Project, 2016), which has been studied extensively and serves as an excellent source of crowdsourced forecasting tournament data akin to the other platforms cited above (Ungar et al., 2012).

While regularized logistic regression was the most effective machine learning technique studied by Shinitzky et al. (2024), its assumptions may not hold in a dataset with a greater variety and number of participants, forecasts, and variables unavailable in the 'cold start' situation Shinitzky et al. considered. Additionally, Shinitzky et al. tested a Bernouli naïve Bayes classifier, but did not examine Bayesian

networks (BNs) with more complex conditional dependencies between predictor variables. Bayesian networks make fewer parametric assumptions than logistic regression and their ability to represent complex conditional dependencies between variables may prove beneficial in modeling more complex forecasting data. Finally, Shinitzky et al. used an unweighted mean as their aggregation procedure in their baseline predictor variable. More sophisticated aggregation procedures, such as those that incorporate recalibration and weighting by participant characteristics would likely make the models more predictive. Recalibrated weighted forecasts are also highly accurate predictors on their own, so it would be insightful to compare other modeling approaches to these techniques (Atanasov et al., 2017).

4.4.1 Bayesian Networks to Predict Geopolitical Outcomes

The current study explores Bayesian networks (BNs) as an alternative machine learning approach to improving the predictive power of geopolitical forecasts and compares their relative merits to logistic regression and recalibrated aggregate forecasts. Bayesian networks represent relationships between variables as a graph where connections (also called ‘arcs’) between variables (also called ‘nodes’) represent conditional probabilistic dependencies. Given that the connections represent conditional probabilities, for a BN to be fully specified, all connections must be directed (specifying which variable is conditional on which), and there can be no cyclical connections. Thus, BNs are directed acyclic graphs (DAGs). The absence of an arc between nodes represents conditional independence. The conditional probability distribution of a variable is determined by its parent nodes, and a variable is conditionally independent of its non-descendants, given its parents. Arcs can be learned via algorithmic means, via manual specification, or a combination of the two (Kitson et al., 2023).

Three common types of BNs are: 1) discrete BNs, where all variables are categorical and conditional relationships are represented in conditional probability tables, 2) Gaussian BNs, which represent all variables as normally distributed continuous with conditional relationships usually expressed as linear regression equations; and 3) hybrid BNs, which include both discrete and continuous nodes, usually with the constraint that discrete nodes cannot inherit from continuous nodes (Kitson et al., 2023). In the current study, only discrete BNs are used since the binary outcome variable could not inherit from continuous variables. A process of discretizing continuous variables is thus employed to meet the constraints of discrete BNs.

Bayesian networks allow one to model complex conditional dependencies between variables, where the relationships between descendent variables depend on the states of ancestor variables. Learning the pattern of conditional dependencies between predictor variables may improve the overall accuracy of the BNs. For example, divergence from a consensus forecast is usually related to poor performance. However, in some cases where a divergent forecaster is otherwise accurate, their divergent forecast should be given more weight, as it represents unique information (Atanasov and Himmelstein, 2023). Thus, the relationship between a divergent forecast and the outcome may be different depending on the states of other variables. To examine the effect of complex conditional dependencies on the BNs' performance, complex BNs whose structures are learned primarily through algorithmic means (henceforth 'structure-learned BNs') are compared to naïve flat-structured BNs (henceforth 'flat BNs') that assume independence between all predictor variables.

Discrete BNs make fewer parametric assumptions about the data than do logistic regression models, which assume, among other things, that there are no influential outliers (Stoltzfus, 2011). This assumption could be problematic given the presence of 'superforecasters' in the GJP dataset, who exhibit exceptionally high accuracy, and whose forecasts should be given higher weight, rather than transformed or removed to meet this assumption (Atanasov and Himmelstein, 2023). Furthermore, logistic regression assumes that there are linear relationships between the predictor variables and the log odds of the outcome (Stoltzfus, 2011), which may not hold true in a large dataset with a wider range of questions and participants than Shinitzky et al. (2024). This study compares the performance of BNs to regularized logistic regression, which was the highest performing machine learning method explored by Shinitzky et al. (2024).

The current study also incorporates forecast aggregation and recalibration. Shinitzky et al. (2024) used the absolute distance from a consensus forecast as their baseline predictor, where the consensus was the unweighted average of forecast probabilities for a question. The current study uses the information available in the GJP dataset to create a consensus aggregate that applies recalibration to an aggregate weighted by participant accuracy, forecast update frequency, and temporal proximity. This follows Atanasov and Himmelstein's (2023) procedure for calculating the consensus forecast for their proxy score. The absolute difference from the recalibrated aggregate is used as the baseline predictor in the current study. The consensus aggregate is also treated as a separate predictive model to serve as a basis of comparison for the other models, since this form of recalibration and aggregation can produce highly accurate forecasts (Atanasov et al., 2017).

Given that the recalibrated weighted aggregate is used as a component of the baseline predictor in all models, the first hypothesis of this study is that as more variables are included in the models, the BN with complex dependencies will have higher performance than the recalibrated weighted aggregate alone. The second hypothesis is that both structure-learned and flat BNs will have a higher performance than the equivalent logistic regression models, since BNs make fewer parametric assumptions than logistic regression, some of which are likely violated in a dataset like the GJP. The third hypothesis is that as the models include more variables, structure-learned BNs will have higher performance than flat BNs, since capturing more complex conditional dependencies will likely increase the information available for predicting the outcome.

4.5 Methods

To describe the methodology of the data analysis, I first characterize the dataset and the transformations applied to derive the continuous predictor variables. Subsequently, I outline the discretization process for the continuous variables, followed by the procedures employed to train the BN and logistic regression models. Finally, I detail the evaluation protocols applied to all models.

The dataset consisted of binary, non-conditional and non-voided questions from all four years of the GJP ACE tournament, yielding 694,442 forecasts made by 9480 forecasters on 303 questions. An example question is “Will Serbia be officially granted EU candidacy by 31 December 2011?”. Each forecast consisted of a probability value (also called ‘forecast value’) associated with the first binary option of the question and included a timestamp of when the forecast was made and when the question closed. Four continuous predictor variables (described in detail below) were derived from the dataset which served as the basis of models predicting the binary outcome of the question (0 or 1) on a per-forecast basis. Depending on the model type, the continuous predictor variables were also discretized following multiple discretization configurations of each variable (described further below). For the sake of clarity, the models are grouped into tiers which correspond to the number of variables included as predictors. Each subsequent tier includes the predictors of previous tiers. Details of the predictor variables are described next according to the tier of model they first appeared in, followed by the discretization process.

The first tier of models included only the absolute difference of the forecast value from a consensus prediction value at the time of the prediction, similar to Shinitzky et al.'s (2024)'s baseline model. First, forecast values of .00 and 1.00 were replaced with .001 and .999, respectively, to avoid distortions in

logarithmic operations using 0 (see Satopää et al. (2014) for a similar procedure). All predictors used this corrected forecast value, including the consensus prediction, described next.

The consensus prediction was a recalibrated weighted aggregate forecast, following the procedure described in Atanasov et al. (2017) and Atanasov and Himmelstein (2023). The weighted aggregate is described in (1),

$$\bar{p}_{t,k,l} = \left(\frac{1}{\sum_{t,i} \{d_{t,i,l} \times w_{t,i,l}\}} \right) \times \sum_{t,i} \{d_{t,i,l} \times w_{t,i,l}\} \times p_{t,i,k,l} \quad (1)$$

where $\bar{p}_{t,k,l}$ is the aggregate across forecasters, i , that depends on time, t , question, l , and outcome, k . The decay value, $d_{t,i,l}$, was set to 1 for 72% of the most recent forecasts within a question, and 0 otherwise. The weight value, $w_{t,i,l}$, depended on the accuracy score of a forecaster, c , and the update frequency score for a forecaster within a question, f (see (2)).

$$w_{t,i,l} = c_{t,i}^{\gamma} \times f_{t,i,l}^{\delta} \quad (2)$$

The accuracy score was based on the cumulative mean standardized Brier score (MSBS) for all questions that had resolved by the time the forecast was made (standardization was within-question). The frequency score was based on the number of forecasts made on the question up to the point of the current forecast. Both the accuracy and frequency score were rescaled to the range [0.1 to 1.0], where the least accurate / least frequent updater would receive 10% of the weight of the most accurate / most frequent updater. Both weight exponents, γ , and δ , were held at 1.0. The accuracy weight was omitted when no questions had been resolved for a participant at the time of the forecast. The weighted aggregate was then recalibrated following (3),

$$\hat{p}_{t,k,l} = \frac{\bar{p}_{t,k,l}^a}{\bar{p}_{t,k,l}^a + (1 - \bar{p}_{t,k,l})^a} \quad (3)$$

where the extremizing constant of a was 1.5.

The weighting and recalibration parameters were the same as those in Atanasov and Himmelstein (2023) and were not optimized for the dataset, so the predictive power of the aggregate forecast should be considered conservative. This recalibrated weighted aggregate was calculated for each forecast, such that it reflected the aggregate at the time the forecast was made.

The second tier of models included the forecast value as an additional predictor. The third tier of models included the number of days prior to the close of the question the forecast was made (henceforth ‘days prior’). The fourth tier of models included the MSBS for a user at the time the forecast was made. For this tier of models, forecasts made by a participant prior to a single resolved question were discarded to avoid temporal leakage of accuracy information, yielding 619, 873 valid forecasts.

The MSBS was calculated by first calculating the binary Brier score of each forecast, i.e., the squared difference between the forecast value and the outcome (0 or 1). The forecast Brier scores were then standardized within each question, where the standardized value was the difference of each forecast value from the mean of the most recent forecasts made by other participants on the question, divided by their population standard deviation. If the standard deviation within a question was zero, the standardized score of a forecast was set to zero, to indicate no difference from the mean. Finally, the standardized Brier scores were averaged across questions that had resolved at the time of the current forecast, such that the MSBS reflected the average Brier score for a participant at the time of the current forecast. This score was then standardized and winsorized to be between -3 and 3 standard deviations of the mean due to the presence of some extreme outliers that negatively impacted the discretization process. It appears that extreme outliers were usually caused by early clustering of forecast values around a single mean value, leading to a small standard deviation, which would yield extremely large standardized scores for the small number of forecasts that were highly divergent from the mean. Given that these outliers appeared to legitimately reflect a large difference from the mean, but their exact value did not validly reflect the extremeness of the scores, winsorization was deemed an appropriate transformation. The discretization process also rendered the exact values of the outliers largely moot, as all extreme scores would tend to be binned into one discrete level.

For each tier of model, a set of discretization configurations for the variables were tested. For each configuration of discretized variables, three models were trained: a structure-learned BN, a flat BN, and a regularized logistic regression model using dummy variables to represent the discrete levels. A regularized

logistic regression model was also trained on the continuous versions of the variables included in each tier of model. The details of each of these methodologies are described after explaining the discretization process. Finally, the raw recalibrated weighted aggregate forecast described in (3) was used as a separate baseline model to which the other models were compared, where the aggregate forecast value was treated as the classification probability.

4.5.1 Discretizing the Variables

The continuous variables described above were discretized using the Hartemink algorithm, as implemented in bnlearn version 5.1 (Hartemink, 2001; Scutari, 2010; Scutari and Silander, 2025). The algorithm attempts to retain as much total pairwise mutual information as possible while determining where to make discrete breaks in continuous variables (where total mutual information is the sum of mutual information between each pair of variables). The algorithm begins by discretizing the continuous variables into many levels (in this study, 120 levels of equal intervals). It then iteratively selects two neighboring levels in each variable to coalesce such that the least amount of total mutual information is lost in the coalescence. At each iteration, the number of discrete levels per variable decreases by one until it reaches either a pre-specified number, or a minimum of 1, where total mutual information = 0.

While choosing the precise number of discrete levels to retain for each variable requires some subjective judgement, the choices were informed by the information presented in Figure 4.1. The figure depicts the pairwise mutual information of each variable when all variables were discretized into the same number of levels, starting from 120 levels of equal intervals. A few trends can be observed. First, some variables contained much more pairwise information than others, with forecast value and difference from the aggregate containing roughly 43% and 32% of the information, respectively. In comparison, days prior and MSBS contained only about 8% and 4%, respectively. Second, the information decreased roughly logarithmically with each fewer level, with some variables showing a sharper decline as they approached the minimum of 2 levels. Notably, forecast value showed a sharp decline below 6 levels, and difference from the aggregate showed a sharp decline below 4 levels. Friedman et al. (2018) suggest that the average number of discrete bins required to avoid significant declines in precision for the GJP competition was 6 (11 for superforecasters), which additionally informed how many discrete levels the forecast value variable was given.

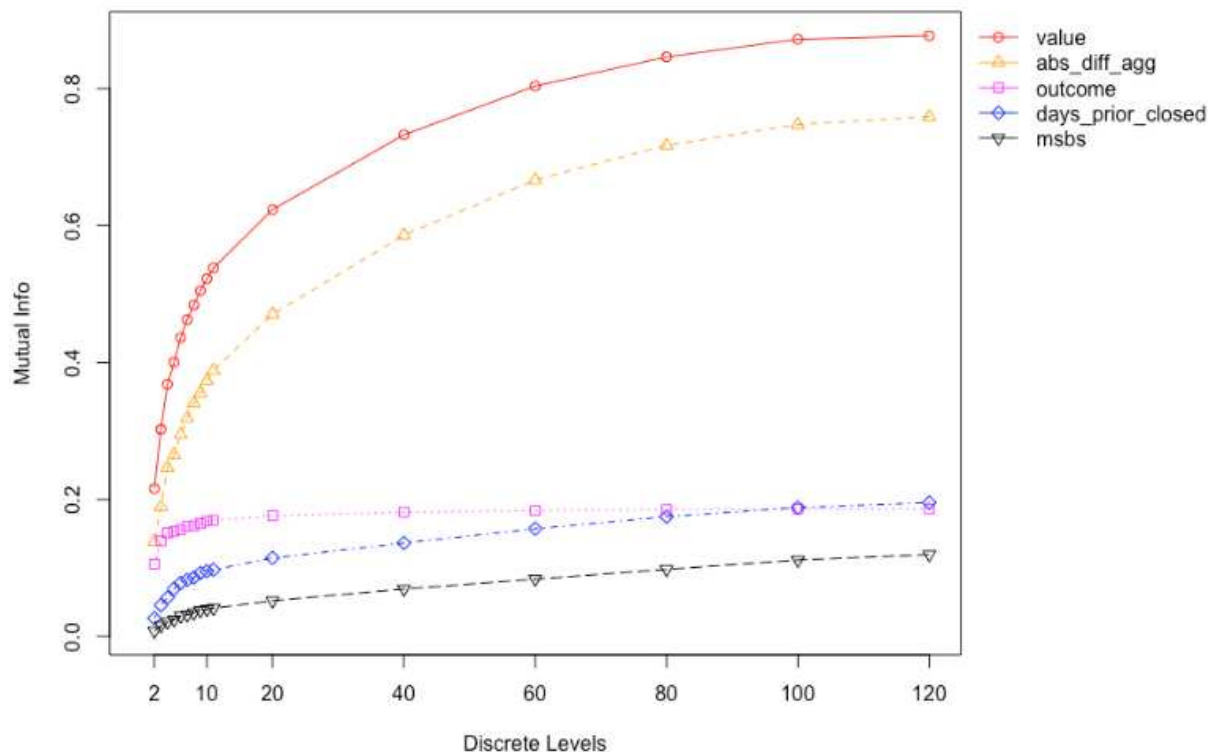


Figure 4.1. Total mutual information as a function of the number of discrete levels

Note. Abs Diff Agg = Absolute Difference from the Aggregate, MSBS = Mean Standardized Brier Score.

These observations resulted in a few heuristics that were flexibly applied in deciding how many discrete levels each variable would receive. The first heuristic was that more levels should be given to variables that contained more mutual information. The second heuristic was that forecast value would have a target of between 11 and 6 bins, and absolute difference from the aggregate was given a target of at least 4.

As discussed by Hartemink (2001), there exists a trade-off between using more discrete levels per variable, which retains more mutual information, and the complexity of the resulting model, which can yield more parent-child relationships and thus more cells in the conditional probability tables of the BNs that must be estimated. More cells lead to fewer observations per cell, and nodes that are rarely or never observed cannot be estimated with certainty. Prior probabilities can be used to estimate unobserved combinations, as was done in this study, but in the absence of strong theoretical underpinnings of the prior specifications, overreliance on priors risks a poorly specified model.

To avoid over-reliance on priors in the BN parameterization caused by rarely observed parent-child configurations, discretization configurations were subject to the hard constraint that no more than 10% of the combinations of all discrete levels in the configuration could be unobserved. Following these heuristics and constraints, a set of discretization configurations was created for each tier of models (included in Table 4.1).

4.5.2 Training the Bayesian Networks

Training a BN is a two-step process of first learning the structure of the conditional dependencies between variables, followed by fitting the conditional probability tables of the variables, given the structure and the data. Given that the models were evaluated using 10-fold cross-validation (see below for details), 10 separate BN structures were learned on the training portion of each fold of the cross-validation setup. Structures were learned using the PC-stable algorithm implemented in bnlearn with the mutual information test. Arcs from the outcome to predictor variables were forbidden to ensure the predictors were parents of the outcome. Directed connections between variables were added if they appeared in the learned structures of $> 6/10$ of the training folds. Undirected connections with $> 6/10$ support were added and manually directed based on theoretical considerations. The same structure was used in each fold during the parameter fitting and cross-validation phase. The parameters of both the structure-learned and flat BNs were fit using maximum Bayesian posterior estimation, with a uniform prior corresponding to the Bayesian Dirichlet equivalent score (Scutari, 2025). Probability predictions about the outcome were generated from Bayesian likelihood weighting using the default of 500 random samples.

4.5.3 Logistic Regression Models

Logistic regression is a statistical methodology that models the probability of an outcome variable ($\hat{Y}$) being in one category or another (1 or 0) using a linear regression equation where each predictor variable (X_i) is weighted by its respective β coefficient. The β coefficients represent the slope of the regression line (which quantifies how much the outcome variable changes with a one-unit change in the predictor variable) and reveal the unique contribution of each predictor variable, after adjusting for the others (Stoltzfus, 2011). Regularization techniques apply a penalty to the β coefficients according to a tuning parameter, λ , which specifies the strength of the penalty. Regularization techniques are useful for logistic regression models with many potentially correlated predictor variables. Lasso/L1 regularization tends to pick one correlated coefficient and shrink the rest to zero, whereas ridge/L2 regularization tends to shrink correlated coefficients towards each other (Friedman et al., 2010).

The regularized logistic regression models in this study were fit with the glmnet R package version 4.1.10 (Friedman et al., 2010, 2025). For cross-validation, glmnet uses a grid search to find the value of λ that minimizes the mean loss across the 10 folds. This λ is then used in all 10 folds of the subsequent evaluation of the model. Following Shinitzky et al. (2024), the Tier 1 discretized model was fit with ridge/L2 regularization. Lasso/L1 regularization was used for all other tiers. Because the Tier 1 continuous model used only a single predictor, regularization using glmnet was inappropriate, so non-regularized logistic regression was performed using the glm function from the stats package in R (version 4.5.0).

4.5.4 Model Evaluation

Ten-fold cross validation was used to evaluate the models. The folds were balanced across outcome, with questions being segregated across folds to avoid information leakage. The same folds were used for each model within a discretization configuration, whereas folds were randomly generated across configurations. All models were evaluated using AUC as calculated by the pROC R package, version 1.19.0.1 (Robin et al., 2011).

4.6 Results

The AUC results suggest that the baseline recalibrated aggregate forecast was superior to both the Bayesian and logistic regression approaches, contrary to the first hypothesis of this study (see Table 4.1 for AUC results for all models). Both Bayesian approaches showed higher AUC scores than both logistic regression approaches, consistent with the second hypothesis of the study. However, the structure-learned BNs showed equivalent AUC scores to the flat BNs, contrary to the third hypothesis of the study. The discretized and continuous regularized logistic regression approaches both showed roughly equivalent AUC scores across all model configurations.

Table 4.1. Performance Metrics for Baseline, Bayesian Network, and Logistic Regression Models

Config	No. discrete levels per variable				Mean (SD) AUC				BIC		
	Abs diff agg	Fcast Value	Days Prior	MSBS	BN	Flat BN	Discrete LogReg	Continuous LogReg	Aggregate	BN	Flat BN
Baseline									.985 (.018)		
Tier 1											
1.0	80	0	0	0	.602 (.047)	.602 (.047)	.606 (.047)	.580 (.051)		-3013712	-3013712
Tier 2											
2.0	4	10	0	0	.939 (.056)	.939 (.056)	.830 (.071)	.825 (.074)		-2334986	-2501996
2.1	5	8	0	0	.938 (.041)	.938 (.041)	.831 (.056)	.824 (.061)		-2444450	-2626347
2.2	5	7	0	0	.939 (.036)	.939 (.036)	.831 (.050)	.826 (.046)		-2348550	-2525157
2.3	4	6	0	0	.940 (.036)	.940 (.036)	.839 (.060)	.834 (.060)		-1975226	-2128432
Tier 3											
3.0	3	6	3	0	.925 (.046)	.925 (.045)	.827 (.064)	.814 (.059)		-2613377	-2785402
3.1	3	5	3	0	.932 (.039)	.932 (.039)	.829 (.040)	.829 (.032)		-2455466	-2622747
3.2	3	5	2	0	.929 (.049)	.929 (.049)	.817 (.078)	.829 (.072)		-2143187	-2300066
Tier 4											
4.0	3	4	2	2	.929 (.046)	.930 (.046)	.824 (.059)	.833 (.056)		-2157799	-2297388

Note. Abs diff agg = Absolute Difference from Aggregate, AUC = Area Under the (receiver operating characteristic) Curve, BIC = Bayesian Information Criterion, BN = Bayesian Network (structure-learned), Fcast = Forecast, Flat BN = Flat Bayesian Network, LogReg = Logistic Regression, MSBS = Mean Standardized Brier Score.

For the Bayesian and logistic regression models, the highest AUC values were found when the models included only the forecast value and the absolute difference from the aggregate. Including days prior and MSBS did not improve AUC scores. Interestingly, within the best-performing tier of models (Tier 2), the number of discrete levels for both variables did not appear to have much effect on AUC. Descriptively, the highest performing model had the fewest discrete levels for forecast value and absolute difference from the aggregate (six and four levels, respectively). It thus appears that six discrete levels for forecast value was sufficient to avoid errors due to imprecision, as might be suggested by Friedman et al. (2018).

To further investigate why the structure-learned and flat BNs did not differ in AUC, the Bayesian Information Criterion (BIC) score of the models using the entire dataset were recorded (see final two columns of Table 4.1). The BIC score measures the log-likelihood of the data given the model, minus a penalty for model complexity (Kitson et al., 2023). This can serve as an indicator of whether any gains in explanatory power were worth the cost of a more complex model (e.g. with more structure / conditional dependencies). More positive values (closer to zero) represent a more likely and/or more parsimonious model. The structure-learned BNs had consistently more positive BIC scores than flat BNs, indicating that there were gains in explanatory power that overcame the penalty for greater complexity. However, it appears these gains were not large enough to result in meaningfully more predictive models. The most positive BIC score belonged to the configuration with the highest AUC and included only forecast value and absolute difference from the aggregate, with the fewest discrete levels each.

4.6.1 Relationships Between Predictor Variables

Examining the arrangement of the structure-learned BN models can show how the variables related to one another and can reveal information about the efficacy of the model. The structure-learning algorithm found that the outcome variable was conditionally dependent on absolute difference from the aggregate, forecast value, and days prior, for all models that included these variables. The Tier 4 model found that MSBS was conditionally independent of outcome, given the other predictor variables (see Figure 4.2). These results suggest that MSBS did not provide unique information related to the outcome beyond what was already present in the other variables. Overall, the structures of the BNs were consistent across models, and resembled that of Figure 4.2 (potentially with fewer variables, depending on the model tier).

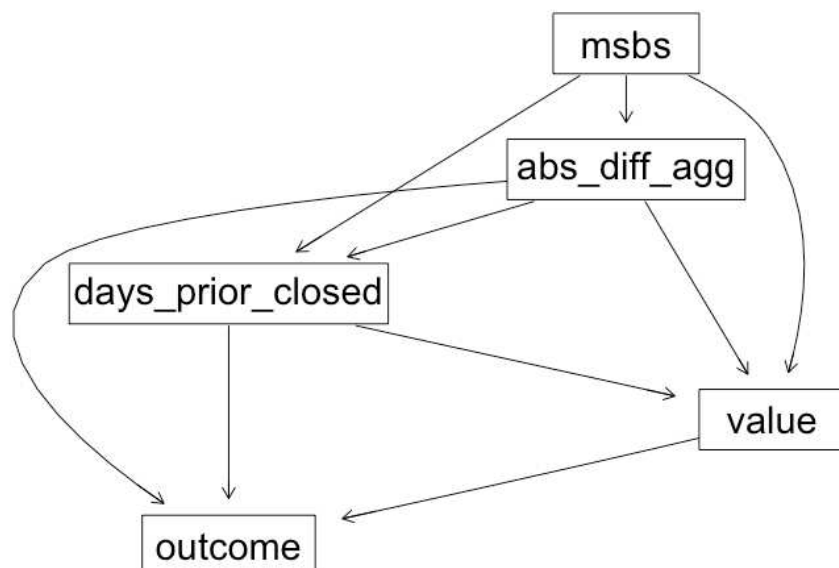


Figure 4.2. The Bayesian network structure from Tier 4

Note. Abs Diff Agg = Absolute Difference from the Aggregate, MSBS = Mean Standardized Brier Score.

In a similar vein, examining the coefficients from the logistic regression models can shed light on the how the variables were related and the efficacy of the models. The coefficients from the Tier 4 continuous logistic regression model suggested that MSBS and absolute difference from the aggregate contributed little unique predictive power, as their coefficients shrank to zero from the lasso regularization (see Table 4.2). Forecast value was the variable with by far the strongest relationship to the outcome. These results also suggest that the predictor variables contained redundant information for predicting the outcome.

Table 4.2. Coefficients from the Tier 4 Lasso Logistic Regression Model with Continuous Variables

Predictor	Coefficient
(Intercept)	-2.33
Forecast Value	2.61
Days Prior Closed	-0.0003
Abs Diff Agg	-
MSBS	-

Note. Abs Diff Agg = Absolute Difference from the Aggregate, MSBS = Mean Standardized Brier Score.

The results from the discretized Tier 4 logistic regression model are more complex but suggest that the most influential variables were forecast value and large values of absolute difference from the aggregate (see Table C1). MSBS and long-range forecasts received relatively little weight.

4.6.2 Testing the Assumptions of Logistic Regression

Examining the assumptions of the logistic regression models can illuminate why they had lower AUC values than the Bayesian models. The following five assumptions of logistic regression were tested on a non-regularized version of the Tier 4 continuous logistic regression model using the full dataset: independence of errors, linearity of predictor variables and the log-odds of the outcome, lack of multicollinearity among predictors, a lack of influential outliers, and a sufficiently large sample size (Stoltzfus, 2011). The assumption of independence of errors was violated on a priori grounds, since participants could provide repeated forecasts on the same question, leading to non-independence due to participant and question. Violating this assumption tends to artificially reduce the variance in parameter estimates, yielding overly small confidence intervals and p values, and can also bias parameter estimates (Ives and Zhu, 2006; Roberts et al., 2017). It should be noted that BNs also typically assume independence of errors, so this is a shortcoming that applies to both types of models tested in this study (Azzimonti et al., 2022).

The assumption of a linear relationship between the predictors and the log-odds of the outcome and was tested with the Box-Tidwell test, where a logistic regression model was fit to the original predictors and interaction terms consisting of those predictors multiplied by their natural logarithms. All interaction terms were statistically significant ($p < .001$) and had large absolute z values (> 17), indicating a violation of the assumption of linearity. Likewise, visual inspection of scatterplots of predictors versus the log odds of the outcome suggested a violation of linearity in the absolute difference from aggregate variable, as it displays a curvilinear U-shaped pattern (see Figure 4.3). Note that this assumption only applies to

continuous variables and not dummy-coded variables, so it was not violated in the dichotomous logistic regression models (Stoltzfus, 2011).

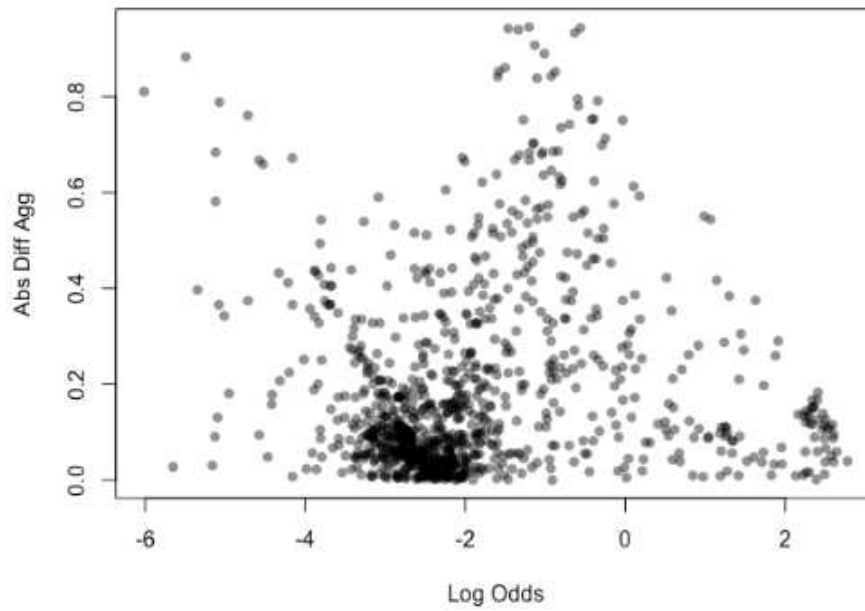


Figure 4.3. Scatterplot of Absolute Difference from Aggregate and Log Odds of Outcome.

Note. To aid visualisation, points are from a random subset of $n = 1000$ data points. Abs Diff Agg = Absolute Difference from the Aggregate.

The assumption of the absence of influential outliers was tested by examining Cook's distances and standardized residuals. The results suggested that approximately 6% of the data had a Cook's distance $> 4/n$, and approximately 1% had an absolute standardized residual > 3 , indicating a mild violation of this assumption. The assumption of the absence of multicollinearity was tested by examining the variance inflation factors (VIF) of the predictor variables. All VIF values were < 5 , indicating no violation of this assumption. Finally, the assumption of a sufficiently large sample size was met, since a common rule of thumb is that there should be at least 10 observations of the least common outcome per independent variable, and the current sample consisted of $> 600,000$ observations with the least common outcome being approximately 20% (Stoltzfus, 2011). In sum, it appears that some of the assumptions of logistic regression were violated, particularly the assumptions of independence of errors and linearity. This likely

led to somewhat poorly specified models, reducing their predictive ability relative to BNs which make fewer parametric assumptions.

4.7 Discussion

The results suggested that the recalibrated aggregate predictor alone was the most accurate at predicting outcome as measured by AUC, followed by the BN models, then regularized logistic regression models. Structure-learned BNs were not superior to flat BNs. Discretizing the variables did not have a meaningful effect on the performance of the logistic regression models.

There are a few plausible explanations for why BNs did not out-perform the recalibrated aggregate. The first is that the aggregate included information redundant with the other variables included in the models, namely, time information, forecast value, and accuracy. It was believed that absolute difference from the aggregate, conditional on the other variables, such as forecast value and MSBS, would provide a unique signal that would improve AUC over the raw aggregate. However, it seems that while conditional dependencies were found by the structure learning algorithm, they did not add enough additional information to increase AUC over flat networks. The fact that MSBS was not independently related to outcome, as found by the Tier 4 structure-learned BN, provides additional evidence that the variables contained redundant information related to participant accuracy.

Another plausible explanation for why the Bayesian methods did not outperform the raw aggregate is that the discretization process removed too much information from the continuous variables. As can be seen in Figure 4.1, reducing the number of discrete levels in forecast value from 120 to 10 results in a loss of approximately 40% of its total pairwise information. Likewise, the absolute difference from aggregate lost 65% of its total pairwise information between 120 and 5 levels (its highest level beyond Tier 1). Thus, the discretization process may have substantially reduced the information available to the models and thus lowered AUC scores.

The results highlight trade-offs in using discrete BNs to model realistic forecasting data. The upsides are that discrete BNs make few parametric assumptions and can model discrete outcomes directly (unlike typical hybrid BNs, where discrete nodes cannot be the child of continuous nodes). They are also transparent models of conditional dependencies between variables, which can be theoretically illuminating and sometimes highly related to model accuracy. Bayes nets are also able to incorporate

theory-driven information in the form of expert-derived directional dependencies and prior conditional probabilities (Kitson et al., 2023).

The downsides are that the discretization process can substantially reduce the information available in continuous variables. The discretization process also requires a large dataset, as a greater number of discretized levels introduces more cells of conditional probability tables that must be estimated, reducing the number of observations available per cell. This puts a ceiling on the number of variables that can be included in a model. Finally, the discretization process requires some subjective judgement in deciding how many discrete levels each variable should have, with no guarantee of an optimized solution.

The relatively poor performance of the logistic regression models is likely owing to violations of the assumptions of the models, particularly independence of errors and linearity. Transformations of the variables to make them better fit the assumptions of logistic regression and forms of logistic regression better suited to repeated measures / correlated data would likely be an effective solutions to these issues (Stoltzfus, 2011). Non-standard hierarchical models that account for correlated data are available for both logistic regression and BN models, so exploration of these more advanced techniques is warranted, as are more complex blocking designs for the cross-validation procedure to account for dependencies in the data (Azzimonti et al., 2022; Ives and Zhu, 2006; Roberts et al., 2017). Additionally, explorations of interaction effects in the logistic regression models would also likely yield gains like those of conditional dependencies between nodes in a BN.

Future research wishing to expand on this study would first do well to examine a wider range of variables related to the outcome but not strongly related to the recalibrated aggregate. Atanasov and Himmelstein (2023) provide a set of variables worth exploring in this regard. Second, while discretizing all variables is required to model discrete outcomes directly using discrete BNs, a hybrid BN including both discretized and continuous variables could predict a continuous proxy of outcome (such as prediction error), which could then be used as a means of predicting outcomes. One example of this approach is similar to that suggested by Shinitzky et al. (2024), where the prediction probabilities of outcome from a logistic regression model were used as weights in a wisdom-of-the-crowds aggregation approach. Another related approach would be to use a continuous output of a BN as an input to a logistic regression model, which can be more effective than either method used alone (Wei and Dong, 2025). A hybrid BN could use discretized variables that do not fit the assumptions of linear or logistic regression and retain continuous

variables that fit the assumptions of linear regression. A logistic regression model could then use this output, possibly in combination with other non-normal variables that fit the assumptions of logistic regression, to increase the predictive power of the ensemble.

4.8 Conclusions

Overall, the purpose of this study was to explore the utility of BNs in improving geopolitical forecasts, in comparison to logistic regression and recalibration and aggregation. Neither BNs nor logistic regression alone proved superior to state-of-the-art recalibration and aggregation methods under the conditions of this study. The results suggest that the findings of Shinitzky et al. (2024) are not robust to different datasets with different predictor variables, in that logistic regression was not the most accurate machine learning technique of the current study. While the current study did not improve on the accuracy of aggregation and recalibration techniques, the merits of BNs deserve further exploration. Investigations into the use of other predictor variables, data transformations, the effect of discretization, and more advanced forms of BNs (and logistic regression), such as those that can account for correlated data, would be fruitful areas of future research. Furthermore, using ensemble methods that play to the strengths of each methodology remains a promising next step for this research direction.

4.9 Supplementary Materials

Supplementary materials, including analysis code can be found in the following repository:
<https://osf.io/v4hjz/>

4.10 Competing Interests Statement

Competing interests: The author declares none.

4.11 Financial Support Statement

This work was supported by the Fonds de recherche du Québec – Nature et technologies under Grant #303239. The funding agency had no role in the study design, writing of the report, or submission decisions.

4.12 Acknowledgements

The author wishes to thank Serge Robert, Roger Villemaire, and Stephen Martin for their insightful comments and help in editing the final manuscript.

4.13 Declaration of Generative AI and AI-assisted Technologies in the Manuscript Preparation Process

During the preparation of this work the author used Claude Sonnet 3.5 – 4.5 and Claude Opus 3.5 – 4.5 to generate some of the analysis code and help edit the manuscript. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

CHAPTER V

General Discussion

The objective of this thesis is to develop and test cognitive and computational tools to improve rational deliberation and decision-making under uncertainty. This implies ensuring better adherence to epistemic and institutional norms of good reasoning, particularly the norms of probability theory. The primary area of application for this research is the intelligence community (IC), although the insights gained could be applied in other domains like business strategy (Hernandez, 2017). To this end, three projects were completed. Chapter 2 assessed which analytic techniques used in intelligence products predicted better adherence to IC norms and greater objective accuracy. Chapter 3 tested the efficacy of a calibration training paradigm based on a proper scoring rule designed to be easily interpretable and facilitate learning. Chapter 4 used Bayesian networks (BNs), logistic regression, and recalibration and aggregation methods to improve the accuracy of geopolitical forecasts from the Good Judgement Project (GJP). The key findings of the three projects are summarized below, followed by their interpretation and implications, their limitations, and finally suggestions for future research.

5.1 Summary of Results

The study in Chapter 2 used content analysis to code for the use of structured analytic techniques (SATs) and other methods in intelligence products written by teams of intelligence analysts and members of the public during two experiments that attempted to replicate the conditions of the IC. The analyses revealed that adherence to IC norms was positively related to objective accuracy. The techniques of Standardizing Probability Ranges and Timeline were predictive of better adherence to IC norms. However, no specific technique was robustly predictive of accuracy, although modest but inconclusive evidence suggested that Timeline may be also positively related to it. Finally, the number of unique techniques used in an analytic product was positively related to adherence to IC norms, but not accuracy. The lack of conclusive results for accuracy is likely owing to low power, as fewer reports could be scored for this outcome metric. All correlations between technique use and IC norm adherence or accuracy were small, indicating that the use of any singular technique was not a strong signal of either metric.

In Chapter 3, two short-duration training experiments were conducted to improve calibration of subjective confidence using performance feedback that included the Practical scoring rule, the Bias

score, and the Bias score's interpretation. Both experiments found that training with performance feedback did not improve calibration, contrary to the only other known experiment to test the Practical scoring rule in calibration training (Gruetzmacher et al., 2023). Training with outcome feedback also did not improve calibration. As predicted, Actively Open-Minded Thinking (AOMT) was positively related to overall calibration but did not predict change in calibration across time.

In Chapter 4, BNs, logistic regression, and recalibration and aggregation were used to predict the outcomes of geopolitical events using human forecasting data from the GJP. Two types of BNs were tested: flat BNs, where predictor variables were assumed to be independent, and structure-learned BNs, where the conditional dependencies between predictor variables were learned algorithmically. A recalibrated weighted aggregate forecast following the methodology of [Atanasov et al. \(2017\)](#) was used as both a separate predictive model and as a component of the BN and logistic models (the absolute difference of the current forecast from the aggregate was the predictor in these models). The recalibrated aggregate was the most accurate predictor of geopolitical outcomes, followed by both types of BNs, then logistic regression. Structure-learned BNs were not meaningfully more accurate than flat BNs. However, structure-learned BNs had more positive Bayesian information criterion (BIC) scores, indicating the relationships between predictors added information over flat networks. Bayesian networks may have suffered from a loss of information due to the discretization process required for standard discrete BNs. Logistic regression models may have been negatively affected by a violation of the assumption of a linear relationship between the absolute difference from the aggregate and the outcome.

5.2 Interpretation and Implications

Interpretations of the three projects follow, particularly in light of the objective of the thesis to develop and test tools to improve reasoning under uncertainty, and considering potential implications for the IC and broader research literature. Of the over 20 techniques examined in Chapter 2, only Standardizing Probability Ranges and Timeline were robustly predictive of IC norm adherence, and none were robustly predictive of accuracy. The relationship between Timelines and IC norm adherence is surprising and warrants further investigation to determine if this finding replicates in different contexts, and if so, what the mechanisms behind it might be. Likewise, the modest, inconclusive evidence in favour of a positive relationship between Timelines and accuracy is intriguing. Timeline as an analytic technique in the IC has not been the subject of much research, with [Stromer-Galley et al. \(2021\)](#) being a rare example, so this

finding may help point towards fruitful follow-up research. The relationship between Standardizing Probability Ranges and IC norm adherence is less surprising, given that using such probability ranges is explicitly part of the IC norms according to which participants were scored.

The finding that IC norm adherence correlates only moderately with accuracy (zero-order Pearson's $r = .30$) highlights the fact that these capture two distinct assessments of quality of reasoning.

Conceptually, there should be a moderate positive correlation between accuracy and IC norm adherence (i.e., not negative, zero, or perfectly positively correlated), given that accuracy was only one of the eight IC norms evaluated in the SWARM studies, and the other seven should be expected to be imperfectly correlated with accuracy. This said, the study in Chapter 2 is one of very few examples of published research that provides evidence about the validity of the IC rating scale, given that the SWARM reports are the only publicly available intelligence products that have been scored according to a version of the scale (Thorburn et al., 2022). The fact that a report can be highly inaccurate, or not follow logical argumentation, yet still receive a moderately high overall score suggests to some authors that the scale is internally invalid (Thorburn & Gelder, 2021).

One proposed alternative to the current IC rating scale is to have a rule-based aggregation approach, where a high score on some dimensions is necessary for a good overall score. For example, reports would be required to score high on the “Uses clear and logical argumentation” dimension to achieve a high total score (Thorburn et al., 2022). Alternatively, [Ashley Barnett et al. \(2021\)](#) have proposed a different set of norms to measure analytic rigor that was developed through a systematic literature review, an expert panel of academics and IC practitioners, and a survey of Australian intelligence analysts. The authors suggest that analytic rigor is work that is appropriately Logical, Objective, Thorough, Stringent, and Acute (LOTSAs). The LOTSA norm scores correlated modestly with the IC rating scale scores when a subset of SWARM reports were rated with both systems (Spearman's $r = .42$; [Thorburn et al., 2022](#)). Ultimately, questions of the internal validity of the IC rating scale would be better addressed with a larger and more externally valid sample of intelligence products, as discussed further below. While the findings of Chapter 2 were modest with respect to evaluating cognitive tools to aid reasoning under uncertainty, as only a few findings were statistically robust, the study has value as a methodology and as a guide to future research that can investigate similar questions with a different (and ideally larger) sample.

The experiments in Chapter 3 found that neither performance feedback in the form of the Practical score and Bias score, nor outcome feedback had a positive effect on calibration. Plausible reasons for the null effect include possible lack of participant motivation or understanding of the task, and low variability in task difficulty. However, in the broader calibration training literature, multiple null findings (e.g., Du et al., 2012; Sieck & Arkes, 2005; Stone et al., 2023), and even backfire effects (Kelly & Mandel, 2024), have been reported. The results of Chapter 3 cast some doubt on the replicability of the positive result Gruetzemacher et al. (2023) found, which showed that training using the Practical scoring rule on general knowledge questions had positive transfer effects to a forecasting task. The finding is doubtful partly in light of some of the statistical choices those authors made. Specifically, Gruetzemacher et al. (2023) used a post hoc statistical technique intended for observational studies to balance the control and treatment groups on some demographic and cognitive variables, despite the use of random assignment in their experiment, which should have been sufficient to balance the groups. The study in Chapter 3 is one of few to report a positive relationship between AOMT and calibration (along with Haran et al. (2013) and Mellers et al. (2015)), so this adds to the body of literature examining this effect. Regarding the objectives of the current thesis, the study in Chapter 3 serves as a critique of Gruetzemacher et al. (2023) and highlights the apparent fact that calibration training interventions require the presence of many poorly understood contextual variables to be successful. More work in this domain should be done to determine the conditions under which calibration training is robustly effective, as discussed further below.

With respect to the objective of evaluating computational tools that effectively improve judgement and decision-making under uncertainty, the results of Chapter 4 were perhaps the most promising. The results highlight how accurate existing recalibration and aggregation techniques already are. Even a non-optimized version of the procedure of Atanasov et al. (2017) had an area under the receiver operator characteristic (AUC) curve near ceiling (AUC = .985). While BNs and logistic regression were not more accurate than the recalibrated weighted aggregate, there are likely ways to further increase their accuracy with refinements to the models and predictor variables, as discussed below.

The effectiveness of aggregating crowd-sourced forecasts would lend some support to the idea behind the Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE) program, in that crowdsourcing analysis likely increases its overall quality and accuracy, particularly if the outputs of the analysis can be quantified and aggregated. It also lends support to arguments that intelligence analysts

should use precise numeric probability estimates where possible (e.g., Barnes, 2016). This would allow for probability estimates to be aggregated and recalibrated in appropriate situations, potentially using accuracy or other metrics such as coherence with probability theory as aggregation weights (Collins et al., 2024). It would also allow for estimates of accuracy and calibration that would be beneficial for feedback and training for analysts. Finally, precise probability estimates could also be used to assess coherence with probability theory, where incoherent estimates would be a sign of poor analysis. If precise probability estimates are too impractical to implement, then numeric probability ranges could be a substitute, particularly if they are paired with verbal expressions that reflect commonly held understandings of those ranges (Barnes, 2016).

5.3 Limitations

The studies conducted in this thesis have several limitations. Four limitations of the study in Chapter 2 present themselves. First, the methodology was observational, so inferences about causality cannot be drawn. Thus, it cannot be said that using Timelines caused the reports to receive higher adherence to IC norms, for example. Second, the reliability of the codes included in the regression analyses was only enough to draw tentative conclusions (Krippendorff's $\alpha > .667$; [Krippendorff \(2004\)](#)), so the implications of the study should be interpreted as such. Third, the experiments were not a perfect replication of the conditions of the IC, with different time pressures, team compositions, and problem types, among other factors, than would be typical in IC analysis. Fourth, many of the reports were written by untrained members of the public, who nevertheless produced reports of higher quality and accuracy than the professional analysts (likely due to greater total time spent on the reports). These last three points limit the external validity and thus generalizability of the findings. The regression and correlational analyses adjusted for team type, but strong inferences about IC practice should not be made given the composition of teams and the influence of team type on the regression results, as well as the other limitations noted above.

The experiments in Chapter 3 found null results, despite best practices being used to ensure data quality from the online panel participants. It is plausible that these participants were not properly motivated to improve their calibration over the course of the experiment or that the feedback they received was not sufficiently informative for them to improve their calibration over time.

The models tested in Chapter 4 did not account for data correlations from participants forecasting multiple times on the same question and answering different questions than other participants. The cross-validation procedure also only blocked for question and not participant or timing of forecasts. This primarily reduces the generalizability of the results, and possibly introduced some bias (Roberts et al., 2017). The discretization procedure limited the number of variables that could practically be included in the BN models and reduced the total amount of information available in the variables that were included. Finally, while the GJP dataset is relatively representative of other large-scale forecasting platforms, it was somewhat constrained in the subject matter of its questions and the timeframes involved. That is, the subjects were limited to geopolitical questions relevant to the IC and most questions were resolved within a short- to medium-term period of less than a year (median days open = 81.50, IQR = 113). The results of experiments conducted on GJP data may not generalize to more user-driven forecasting platforms such as Metaculus, which host questions about a wide range of topics and has questions that may not resolve for decades.

5.4 Future Research Directions

The results of this thesis suggests some directions for future research. A replication of the study in Chapter 2 using real IC reports would answer questions about the external validity of the original project. A successful replication would put the intriguing finding about Timelines being predictive of both IC norm adherence and (possibly) accuracy on firmer ground. Follow-up research could investigate the potential contexts in which Timelines are most beneficial for analysis and the mechanisms behind it. To avoid the enormous number of person-hours required for both scoring each report using the IC rating system by at least three raters and for a manual content analysis with multiple raters, a semi-automated analysis conducted primarily by large language models (LLMs) could be employed. The IC rating scale scores and content analysis codes used in Chapter 2 could be used in reliability tests of such a replication attempt, where if the new methodology produces results with high agreement with the original findings, there would be greater confidence in the findings derived from a new sample. A separate manual reliability analysis should also be completed on the new sample to confirm the reliability of the methodology.

The experiments conducted in Chapter 3 would benefit from follow-up studies that examine the effects of participant motivation on the efficacy of calibration training, along with probing participants' understanding of the task. One means of manipulating participant motivation would be rewarding participants to varying degrees based on their performance. Once positive training effects are reliably

observed, future studies should also examine the effects of different kinds of pedagogical instructions. Additionally, future studies should investigate more individual difference measures such as the Rational-Experiential Inventory (Pacini & Epstein, 1999) or CRT score (Frederick, 2005) to detect who would benefit most from calibration training. Finally, more research should examine the extent to which these kinds of calibration training exercises transfer to tasks such as forecasting, as well as the longevity of any positive transfer effects.

Future research based on the work in Chapter 4 should consider using models that better account for correlations due to participant and question identity, such as linear mixed models for logistic regression (Ives & Zhu, 2006) or hierarchical Bayesian network models (Azzimonti et al., 2022). Furthermore, a more sophisticated blocking procedure for cross validation should be employed that accounts for correlations due to group and time, as this avoids overfitting and improves the generalizability of results to different samples (Roberts et al., 2017). Future studies could also utilize state-of-the-art BN methodologies that allow for discrete variables to inherit from continuous ones, which would eliminate the need for discretizing all predictor variables (Andrews et al., 2018, 2019). Alternatively, a standard hybrid network could be used to predict a continuous proxy for outcome (such as prediction error) using continuous variables that fit the assumptions of linear regression and a small set of dichotomized variables. This proxy could then be used as a predictor in a separate model that predicts outcomes based on the proxy and other variables. The performance of logistic regression models could be improved by transformations to variables that violate the assumptions of the models. Finally, ensemble techniques that combine the benefits of different modeling paradigms would be a promising area for future research into improving the accuracy of geopolitical forecasts.

CONCLUSION

This thesis examined how reasoning and decision-making could be improved in the intelligence community (IC) in three research projects. The project in Chapter 2 used content analysis of intelligence products to examine which techniques were related to better adherence to IC norms and accuracy. The results showed that greater adherence to IC norms was positively related to accuracy and to the use of Timelines in reports. The project in Chapter 3 used a short-duration training paradigm to improve calibration of subjective judgements on general knowledge questions. While the study failed to show improvements due to training with the Practical scoring rule as performance feedback, it was one of few studies in the literature to demonstrate a positive relationship between actively open-minded thinking (AOMT) and calibration. The project in Chapter 4 used recalibration and weighted aggregation of forecasts from the Good Judgement Project (GJP) in combination with Bayesian networks (BNs) and logistic regression to create more accurate aggregate forecasts. The results showed that the most effective means of improving forecast accuracy was recalibration and aggregation following [Atanasov et al. \(2017\)](#)'s methodology, followed by BNs, then logistic regression.

The thesis provides several contributions to the research literature. The first project is one of very few studies that can shed light on the validity of the IC rating scale and how it relates to analytic techniques used in intelligence products. It also provides a methodological framework for future studies to use and potentially investigate the intriguing finding that Timelines may be particularly beneficial to high quality analysis. The second project likewise provides a methodological framework to investigate the conditions under which short-duration calibration training might be effective. Future research in this area could better explain the discrepant findings of [Gruetzemacher et al. \(2023\)](#). The third project in the thesis found that recalibration and aggregation techniques can yield very accurate probabilistic forecasts, lending support to arguments that probabilities should be used in intelligence analysis where possible.

The thesis also suggests concrete next steps for future research. The results of the first study should be replicated using unclassified intelligence products, potentially leveraging large language models (LLMs) to automate some of the process. Pending replication, experimental studies should be done to determine if there is a causal relationship between using the Timeline technique and higher report quality. Following from the second study, future experiments should examine how participant motivation and understanding play a role in the efficacy of short-duration calibration training. Finally, more advanced machine learning

techniques such as hierarchical BNs or ensemble methods should be explored as means of increasing the accuracy of aggregate probabilistic forecasts.

In sum, the thesis advances the research literature on judgement and decision-making under uncertainty by contributing methodological frameworks, empirical findings that merit follow-up research, and supporting evidence for the continued adoption of probability in intelligence analysis. Theories and methodologies from cognitive and computer science are integrated to achieve better understanding and improvement of human reasoning and decision-making within the IC. The thesis examines three types of tools for improving judgment accuracy and adherence to reasoning norms, including structured analytic techniques, calibration training using an intuitive proper scoring rule, and forecast recalibration and aggregation using machine learning. Probability theory is employed as a normative theory of good reasoning and as a methodological tool throughout the work. The thesis argues for greater adoption of probabilities in IC reasoning, as they reduce ambiguity in communication of uncertainty, allow for formal analysis of coherence with probability theory, serve as the basis of feedback using proper scoring rules, and enable aggregation and recalibration.

APPENDIX A
Example Stimuli from Chapter 3

- A) **Correct!** You were 70% confident.
+ 4.93 points!
- B) **Incorrect!** You were 70% confident.
-7.48 points!

Figure A1. Example trial-level feedback for the performance feedback condition in Experiments 1 and 2

Note. Feedback in the outcome feedback condition consisted of only “Correct! / Incorrect!” indications.

Block number	% Correct	Score (max 100)	Bias	Bias Interpretation
1	60	20.55	0.3	over-confident
2	70	25.65	-0.2	under-confident
3	50	15.85	0.004	well-calibrated

Figure A2. Example block-level feedback for performance feedback condition in Experiments 1 and 2

Note. Feedback in the outcome feedback condition consisted of only block number and percent correct.

APPENDIX B

Dichotomized Logistic Regression Coefficients from Chapter 4

Table B1. Coefficients from the Tier 4 Lasso Logistic Regression Model with Dichotomized Variables

Predictor	Coefficient
(Intercept)	-2.20
Days Prior Closed [0,100.65]	0.91
Days Prior Closed (100.65,549]	-0.00000000027
Forecast Value [0.001,0.208917]	-1.25
Forecast Value (0.208917,0.408517]	-
Forecast Value (0.408517,0.807717]	1.43
Forecast Value (0.807717,0.999]	2.68
Abs Diff Agg [2.58926e-05,0.115111]	-
Abs Diff Agg (0.115111,0.353503]	0.072
Abs Diff Agg (0.353503,0.986473]	-1.18
MSBS [-3,0]	0.014
MSBS (0,3]	-

Note. Abs Diff Agg = Absolute Difference from the Aggregate, MSBS = Mean Standardized Brier Score.

SUPPLEMENTARY MATERIALS A

Supplementary Materials for Chapter 2

Details of the reliability analysis

Seven codes had poor reliability with $\alpha < .667$. Of the codes with poor reliability, High Impact / Low Probability Analysis, Indicators or Signposts of Change, and Key Assumptions Check had very few uses between the two independent coders ($n_{Inde}^{22} < 3$). This made a single disagreement highly detrimental to α and obscures patterns in why reliability might be low. The Other code was underspecified and meant primarily for the purposes of finding new codes to add during development, so it is not surprising that this code showed poor reliability.

For the three remaining unreliable codes (Advanced Statistics and Machine Learning, Node and Link or Tree Diagram (Generic), and Bayesian World View), examining the discrepancies between independent coders revealed that these codes were most likely confused with other structurally or conceptually similar codes. Advanced Statistics and Machine Learning was defined as statistical techniques that do not rely exclusively on the generalized linear model family of regression techniques and included an extensional list of methods which were considered this code. It appears to have been often confused with Mathematical Techniques (Generic), its more generic parent lens, and Statistical Charts or Graphs.

Node and Link or Tree Diagram (Generic) was the generic code for a family of node and link diagram-related codes, including Link Analysis, Bayes Nets, Bayes Tree, and Hypothesis Tree. It appears to have been most often confused with Link Analysis, which designates node and link diagrams that primarily describe relationships between people, locations, events, and pieces of evidence. To a lesser extent, it was also confused with Bayes Nets and Hypothesis Tree.

Bayesian World View required explicit references to Bayesian concepts such as Bayesian reasoning or updating prior probabilities, without using Bayes' Theorem for calculations and without being better described by Reference Class-Informed Forecasting, which involved calculating base rates and updating

²² n_{Inde} = the number of times a code appears in the set of 90 reports coded by two fully trained independent raters.

them in response to context-specific information if deemed relevant. Bayesian World View appears to have been confused with both above techniques due to their conceptual similarity, although low usage of this code ($n_{Inde} = 3$) makes firm conclusions difficult to draw.

A slightly longer or more rigorous training process may have ameliorated some of the low reliability, especially for Advanced Statistics and Machine Learning and Node and Link or Tree Diagram (Generic). This is less clear for the other unreliable codes that had relatively few cases of independent coding, making α highly affected by a few disagreements.

Details of the post-hoc power analysis

To approximate the power of the regression analysis on accuracy, I conducted a post-hoc power analysis using G*Power version 3.1.9.4 (Faul et al., 2009). I examined the power of finding a statistically significant single regression coefficient with a partial correlation as strong as Timeline (partial $R^2 = .06$) in a regression model with 23 predictors, $N = 80$ and $\alpha = .05$. Post-hoc power was .60 and a total sample size of $N = 126$ would be required to achieve power = .80 for this effect size. For the average effect size of the predictors in the model (partial $R^2 = .01$), a total sample size of $N = 780$ would be required.

Table SA1. Descriptions of Content Analysis Codes

Code	Level	Summary Description
Node and Link Diagrams		
Node and Link or Tree Diagram (Generic)	1	A graphic visualization of nodes and links, used to represent generic relationships between variables (including hierarchical relationships as represented by generic Trees).
Link Analysis	2	A graphic visualization of nodes and links which is used to represent primarily social relationships / actors, events, pieces of evidence, or locations.
Hypothesis Tree	2	A tree-like graphical structure is used to examine hypotheses and sub-hypotheses in order better understand the scale of a problem and think through possible alternatives.
Tables		
Table (Generic)	1	A catch-all category for tables used in a report that do not fit the description of other sub-categories of tables.
Evidence Table	2	A system for organizing pieces of evidence in a table, rating their credibility, and giving reasons for their credibility rating.
Analysis of Competing Hypotheses (ACH)	2	A structured analytic technique that uses a grid of hypotheses, evidence, and evaluations of how consistent or inconsistent each piece of evidence is with each hypothesis, in order to determine the most likely hypothesis.
Pros and Cons	2	A structured list (usually as a table) consisting of pros and cons for each option or decision under consideration
Reasoning Stress Test	2	A methodology for identifying flaws in reasoning, rating each flaw's severity, and assessing the overall quality of reasoning in a document.
Other Visualizations		
Timeline	1	A representation of events in chronological order in the form of a graphic, table, or list, but not for the purpose of a time-series graph or table.
Statistical Charts or Graphs	1	A visual (graphical) representation of numerical (quantitative) information, including, but not limited to, line graphs, bar charts, and pie charts
Geographical Map	1	A visual display of a geographic region, often with pertinent details highlighted in some way.
Visual Image (Other)	1	A catch-all category for other visual images that do not fit the description of other lenses
Thinking about People		
Game Theory and Prisoner's Dilemma	1	A formal structure for anticipating the decisions of other actors in a situation and acting in a way to maximize one's own interests
Means Motive Opportunity	1	When determining who is likely responsible for an action, examining whether that person had the means, motive, and opportunity to do so.
Threat Assessment	1	Considering the threat posed by various persons/groups by evaluating their intent and capability of harm.
Reasoning about Probabilities		
Bayesian World View	1	A style of thinking that encourages entertaining multiple hypotheses, considering the background relative probabilities of each hypothesis, and updating those probabilities when new information arrives, but without explicit calculations using Bayes' Theorem.
Reference Class-Informed Forecasting	2	A technique for forecasting where one finds similar past situations to establish a base rate of their occurring (aka 'the outside view'), before considering the fine-

			grained details of the current situation (aka ‘the inside view’) and incorporating the two, if deemed appropriate.
Numeric Probabilities and Odds (Generic)	1		The basic principles of numeric probability, odds, and how to manipulate them. Excludes Bayes’ Theorem and more specific lenses such as the lenses of Bayes Nets, Conditional Probability Calculator, and SWARM Probability Calculator as well as Standardizing Probability Ranges.
Bayes’ Theorem	2		A mathematical formula for calculating the probability of a hypothesis, given some evidence.
SWARM Probability Calculator	3		The SWARM in-built calculator for calculating conditional probabilities and solving probability puzzles.
Conditional Probability Table Calculator	3		An Excel file for calculating ‘common effect’ conditional probabilities, which can be used for filling in the conditional probability tables of Bayes Net software packages like UNBBayes.
Bayes Nets	3		A mathematical means of calculating conditional probabilities between variables using specialized software.
Bayes Tree	2		A tree diagram for solving probability puzzles by breaking a large imaginary group into sub-groups using natural frequencies.
Standardizing Probability Ranges	2		A set lexicon of probability terms are paired with numeric probability ranges in order to facilitate communication and understanding of uncertainty.
Estimating Quantities and Probabilities			
Mathematical Techniques (Generic)	1		A catch-all category for mathematical techniques that are not captured by other lenses, in particular, Three-Point Estimating, Regression Using Generalized Linear Models, Advanced Statistics and Machine Learning, Statistical Charts or Graphs, or Numeric Probability and Odds (Generic) (or one of its sub-categories).
Three-Point Estimating	2		A formula for arriving at an estimate of a quantity or probability by taking the average of a low, most likely, and high estimate.
Regression Using Generalized Linear Models	2		A family of related statistical techniques for determining the line of best fit for one or more predictor variables and an outcome variable
Advanced Statistics and Machine Learning	2		A set of techniques for estimating quantities or probabilities from data that do not rely exclusively on the generalized linear model family of regression techniques.
Structured Analytic Techniques			
Key Assumptions Check	1		A technique for explicitly stating each key assumption of an argument and examining how well supported each assumption is.
Quality of Information Check	1		A semi-structured evaluation of the completeness and soundness of available sources of information
Indicators or Signposts of Change	1		A list of observable events or trends that indicate likely changes to larger, less directly observable, trends or events.
Devil’s Advocacy	1		Challenging a single, strongly held view or consensus by building the best possible case for an alternative explanation.
Team A / Team B	1		Use of separate analytic teams that contrast two (or more) strongly held views or competing hypotheses.
Red Team Analysis	1		An analyst (or group of analysts) attempts to take the point of view of an adversary (often from a foreign country; the Red Team) and presents policies or decisions the adversary would likely take, given their point of view.
High-Impact / Low-Probability Analysis	1		A description of plausible ways an unlikely, yet highly impactful event could happen and possible ways it could unfold

What If? Analysis	1	Assumes that an event has occurred with potential (negative or positive) impact and explains how it might come about.
Brainstorming	1	An unconstrained process for generating new ideas or solutions, usually done in a group.
Alternative Futures Analysis	2	A systematic exploration of possible ways a highly complex and uncertain situation can develop.
Outside-In Thinking	1	A method for considering the key forces that can shape an issue, first by examining the 'external' forces outside of the actor's control, followed by those that the actor can have some influence over ('internal' forces).
Miscellaneous		
ODNI Rating Scale aka IC Rating Scale	1	A rating scale used by the Office of the Director of National Intelligence (ODNI) to evaluate proper analytic tradecraft on a scale of 0 (Poor) to 3 (Excellent) on 8 criteria
Admiralty Code	1	A methodology for rating the reliability of a source on a scale of A to F and the credibility of information on a scale of 1 to 6.
Geolocation	1	A set of techniques for determining the geographical location pictured in a photo or video.
Other	1	A catch-all category for techniques or theoretical frameworks used in reports that are not already captured by the lenses described in the coding manual.
None	1	Where there does not appear to be any technique or theoretical framework used in the report, either described by a lens in the coding manual, nor which would count as an 'Other' lens.

Note. Table spanners represent categories that roughly overlap with the hierarchical structures used to ensure codes were mutually exclusive. Level represents the level of hierarchical relationships between related codes, where higher numbers denote a more specific code. More detailed structure is contained in the full coding manual, a redacted version of which can be found at: osf.io/8n6rz/

Table SA2. Reliability Analysis of All Codes

Code	n_{Full}	n_{Acc}	n_{Inde}	α	95% CI ^a
ACH	10	4	5	0.79	[0.32 - 1.00]
ADC	3	2	0	1.00	[1.00 - 1.00]
ALF	0	0	0	1.00	[1.00 - 1.00]
ASML	7	7	5	-0.02	[-0.05 - 0.00]
BN	9	6	7	1.00	[1.00 - 1.00]
BTH	15	15	8	0.93	[0.74 - 1.00]
BTR	3	3	1	1.00	[1.00 - 1.00]
BRST	0	0	0	1.00	[1.00 - 1.00]
BWV	6	5	3	0.66	[-0.01 - 1.00]
CPTC	2	1	1	1.00	[1.00 - 1.00]
DVAD	0	0	0	1.00	[1.00 - 1.00]
EVTB	12	7	8	0.81	[0.52 - 1.00]
GEOL	18	18	13	0.90	[0.73 - 1.00]
GEOM	52	33	32	0.95	[0.87 - 1.00]
GTPD	2	1	2	1.00	[1.00 - 1.00]
HILP	1	1	1	0.00	[-0.01 - 1.00]
HYTR	4	1	4	0.74	[-0.01 - 1.00]
ISCH	1	1	1	0.00	[-0.01 - 1.00]
KAC	6	4	2	0.56	[-0.02 - 1.00]
LINK	22	14	11	0.75	[0.44 - 0.95]
MMO	10	2	6	1.00	[1.00 - 1.00]
MTG	43	30	25	0.76	[0.59 - 0.90]
NLTG	25	22	15	0.59	[0.35 - 0.80]
NONE	6	0	5	1.00	[1.00 - 1.00]
NPRG	102	64	66	0.79	[0.63 - 0.92]
ODNI	3	3	3	1.00	[1.00 - 1.00]
OITH	0	0	0	1.00	[1.00 - 1.00]
OTHR	18	11	11	0.24	[-0.08 - 0.52]
PRCN	6	0	4	1.00	[1.00 - 1.00]
QIC	0	0	0	1.00	[1.00 - 1.00]
RCIF	17	17	13	0.79	[0.58 - 0.95]
RDTM	0	0	0	1.00	[1.00 - 1.00]
RGLM	10	10	6	0.71	[0.26 - 1.00]
RST	4	4	1	1.00	[1.00 - 1.00]
SPR	29	24	15	0.96	[0.85 - 1.00]
STCG	27	26	19	1.00	[1.00 - 1.00]
SWPC	11	10	8	0.85	[0.55 - 1.00]
TBLG	101	60	64	0.85	[0.71 - 0.95]
THAS	1	1	0	1.00	[1.00 - 1.00]
THPE	4	1	4	0.74	[-0.01 - 1.00]
TMAB	0	0	0	1.00	[1.00 - 1.00]
TMLN	37	23	22	0.83	[0.69 - 0.94]
VSIM	25	18	17	0.71	[0.49 - 0.87]
WHIF	0	0	0	1.00	[1.00 - 1.00]

Note. n_{Full} = number of times a code occurred all 138 reports, n_{Acc} = number of occurrences in the subset of reports that could be coded for accuracy, n_{Inde} = number of occurrences in the non-training subset of 90 reports coded by both independent coders. CI = confidence interval calculated using 1000 bootstrap replications, ACH = Analysis of Competing Hypotheses, ADC = Admiralty Code, ALF = Alternative Futures Analysis, ASML = Advanced Statistics and Machine Learning, BN = Bayes Nets, BTH = Bayes' Theorem, BTR = Bayes Tree, BRST = Brainstorming, BWV = Bayesian World View, CPTC = Conditional Probability Table Calculator, DVAD = Devil's Advocacy, EVTB = Evidence Table, GEOL = Geolocation, GEOM = Geographical Map, GTPD = Game Theory and Prisoner's Dilemma, HILP = High-Impact / Low-Probability Analysis, HYTR = Hypothesis Tree, ISCH = Indicators or Signposts of Change, KAC = Key Assumptions Check, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques

(Generic), NLTG = Node and Link or Tree Diagram (Generic), NONE = None, NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, OITH = Outside-In Thinking, OTHR = Other, PRCN = Pros and Cons, QIC = Quality of Information Check, RCIF = Reference Class-Informed Forecasting, RDTM = Red Team Analysis, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), THAS = Threat Assessment, THPE = Three-Point Estimating, TMAB = Team A / Team B, TMLN = Timeline, VSIM = Visual Image (Other), WHIF = What If? Analysis.

Table SA3. Posterior Summaries of All Coefficients Predicting Quality of Reasoning

Coefficient	M_{β}	SD_{β}	95% CI	$P(incl)$	$P(incl D)$	BF_{incl}
Intercept	0.00	0.08	[-0.16 - 0.15]	1.00	1.00	1.00
Public Team	0.27	0.16	[-0.03 - 0.60]	1.00	1.00	1.00
ACH	0.00	0.08	[0.00 - 0.23]	0.09	0.08	0.87
ADC	0.01	0.14	[-0.14 - 0.16]	0.09	0.08	0.86
BN	0.16	0.28	[-0.01 - 0.82]	0.11	0.31	3.61
BTH	0.01	0.07	[-0.00 - 0.22]	0.09	0.08	0.85
BTR	0.09	0.29	[-0.03 - 0.97]	0.10	0.16	1.72
CPTC	-0.02	0.18	[-0.51 - 0.03]	0.09	0.09	0.91
EVTB	0.25	0.33	[0.00 - 0.91]	0.12	0.46	6.15
GEOL	0.00	0.06	[-0.14 - 0.06]	0.09	0.08	0.87
GEOM	0.01	0.06	[-0.02 - 0.17]	0.10	0.10	1.10
GTPD	-0.01	0.16	[-0.48 - 0.01]	0.09	0.08	0.82
HYTR	0.27	0.44	[-0.01 - 1.25]	0.12	0.35	4.13
LINK	0.05	0.14	[0.00 - 0.43]	0.10	0.20	2.09
MMO	0.03	0.13	[0.00 - 0.41]	0.10	0.12	1.30
MTG	0.06	0.13	[0.00 - 0.41]	0.11	0.23	2.52
NPRG	0.03	0.10	[0.00 - 0.32]	0.10	0.16	1.72
ODNI	0.01	0.14	[-0.23 - 0.10]	0.09	0.08	0.87
PRCN	0.02	0.13	[0.00 - 0.50]	0.10	0.10	1.07
RCIF	0.00	0.07	[-0.12 - 0.01]	0.09	0.08	0.85
RGLM	0.02	0.10	[0.00 - 0.32]	0.10	0.10	1.04
RST	-0.04	0.19	[-0.71 - 0.04]	0.10	0.11	1.18
SPR	0.33	0.27	[0.00 - 0.76]	0.14	0.71	15.51
STCG	0.00	0.06	[0.00 - 0.22]	0.09	0.09	0.92
SWPC	-0.01	0.09	[-0.20 - 0.11]	0.09	0.09	0.93
TBLG	0.08	0.16	[0.00 - 0.49]	0.11	0.29	3.28
THPE	0.01	0.15	[-0.14 - 0.24]	0.09	0.09	0.94
TMLN	0.23	0.22	[0.00 - 0.61]	0.13	0.62	10.73
VSIM	0.00	0.05	[-0.13 - 0.02]	0.09	0.08	0.86

Note. Public Team was included in the null model. M_{β} = mean β coefficient estimate, SD_{β} = standard deviation of β coefficient estimate, $P(incl)$ = prior inclusion probability, $P(incl|D)$ = posterior inclusion probability, BF_{incl} = Bayes Factor for inclusion, ACH = Analysis of Competing Hypotheses, ADC = Admiralty Code, BN = Bayes Nets, BTH = Bayes' Theorem, BTR = Bayes Tree, CPTC = Conditional Probability Table Calculator, EVTB = Evidence Table, GEOL = Geolocation, GEOM = Geographical Map, GTPD = Game Theory and Prisoner's Dilemma, HYTR = Hypothesis Tree, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques (Generic), NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, PRCN = Pros and Cons, RCIF = Reference Class-Informed Forecasting, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), THPE = Three-Point Estimating, TMLN = Timeline, VSIM = Visual Image (Other).

Table SA4. Zero-Order Correlations of All Predictors of Quality of Reasoning

Variable	Pearson's <i>r</i>
Public Team	0.17
ACH	0.09
ADC	0.06
BN	0.18
BTH	0.04
BTR	0.07
CPTC	-0.06
EVTB	0.27
GEOL	0.03
GEOM	0.08
GTPD	-0.05
HYTR	0.22
LINK	0.17
MTG	0.16
MMO	0.15
NPRG	0.19
ODNI	0.00
PRCN	0.06
RCIF	0.01
RGLM	0.07
RST	-0.05
SPR	0.27
STCG	0.08
SWPC	0.00
TBLG	0.24
THPE	0.12
TMLN	0.28
VSIM	0.05

Note. ACH = Analysis of Competing Hypotheses, ADC = Admiralty Code, BN = Bayes Nets, BTH = Bayes' Theorem, BTR = Bayes Tree, CPTC = Conditional Probability Table Calculator, EVTB = Evidence Table, GEOL = Geolocation, GEOM = Geographical Map, GTPD = Game Theory and Prisoner's Dilemma, HYTR = Hypothesis Tree, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques (Generic), NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, PRCN = Pros and Cons, RCIF = Reference Class-Informed Forecasting, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), THPE = Three-Point Estimating, TMLN = Timeline, VSIM = Visual Image (Other).

Table SA5. Posterior Summaries of All Coefficients Predicting Accuracy

Coefficient	M_{β}	SD_{β}	95% CI	$P(incl)$	$P(incl D)$	BF_{incl}
Intercept	0.00	0.10	[-0.20 - 0.22]	1.00	1.00	1.00
Public Team	0.33	0.19	[-0.05 - 0.70]	1.00	1.00	1.00
ACH	-0.17	0.36	[-1.12 - 0.09]	0.19	0.27	1.58
ADC	0.00	0.21	[-0.47 - 0.51]	0.18	0.13	0.72
BN	0.04	0.17	[-0.27 - 0.57]	0.18	0.17	0.89
BTH	0.11	0.21	[0.00 - 0.65]	0.20	0.32	1.91
BTR	0.03	0.18	[-0.22 - 0.66]	0.18	0.14	0.77
EVTB	0.12	0.26	[-0.01 - 0.88]	0.19	0.27	1.56
GEOL	0.02	0.11	[-0.09 - 0.45]	0.17	0.15	0.82
GEOM	0.03	0.11	[-0.01 - 0.40]	0.18	0.18	0.99
LINK	0.02	0.10	[-0.06 - 0.41]	0.18	0.15	0.81
MTG	0.00	0.07	[-0.23 - 0.18]	0.18	0.14	0.73
MMO	0.07	0.27	[-0.14 - 1.03]	0.18	0.17	0.94
NPRG	0.05	0.16	[-0.01 - 0.55]	0.19	0.20	1.14
ODNI	-0.03	0.21	[-0.70 - 0.39]	0.18	0.16	0.85
RCIF	0.03	0.13	[-0.05 - 0.51]	0.18	0.17	0.92
RGLM	0.03	0.12	[-0.20 - 0.37]	0.18	0.15	0.81
RST	0.29	0.46	[-0.03 - 1.31]	0.20	0.38	2.50
SPR	0.01	0.07	[-0.14 - 0.25]	0.18	0.13	0.71
STCG	-0.03	0.12	[-0.44 - 0.00]	0.18	0.17	0.94
SWPC	-0.02	0.12	[-0.43 - 0.08]	0.18	0.15	0.82
TBLG	-0.01	0.08	[-0.32 - 0.10]	0.18	0.14	0.74
TMLN	0.12	0.20	[0.00 - 0.62]	0.20	0.37	2.34
VSIM	0.02	0.10	[-0.09 - 0.36]	0.18	0.16	0.87

Note. Public Team was included in the null model. M_{β} = mean β coefficient estimate, SD_{β} = standard deviation of β coefficient estimate, $P(incl)$ = prior inclusion probability, $P(incl|D)$ = posterior inclusion probability, BF_{incl} = Bayes Factor for inclusion, ACH = Analysis of Competing Hypotheses, ADC = Admiralty Code, BN = Bayes Nets, BTH = Bayes' Theorem, BTR = Bayes Tree, EVTB = Evidence Table, GEOL = Geolocation, GEOM = Geographical Map, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques (Generic), NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, RCIF = Reference Class-Informed Forecasting, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), TMLN = Timeline, VSIM = Visual Image (Other).

Table SA6. Zero-Order Correlations of All Predictors of Accuracy

Variable	Pearson's <i>r</i>
Public Team	0.26
ADC	0.10
ACH	-0.08
BN	0.14
BTH	0.22
BTR	0.03
EVTB	0.19
GEOM	0.10
GEOL	0.07
LINK	0.14
MTG	-0.01
MMO	0.15
NPRG	0.15
ODNI	-0.19
RCIF	0.06
RGLM	0.10
RST	0.19
SPR	0.01
STCG	-0.07
SWPC	-0.06
TBLG	-0.09
TMLN	0.28
VSIM	0.15

Note. ACH = Analysis of Competing Hypotheses, ADC = Admiralty Code, BN = Bayes Nets, BTH = Bayes' Theorem, BTR = Bayes Tree, EVTB = Evidence Table, GEOL = Geolocation, GEOM = Geographical Map, LINK = Link Analysis, MMO = Means Motive Opportunity, MTG = Mathematical Techniques (Generic), NPRG = Numeric Probabilities and Odds (Generic), ODNI = ODNI Rating Scale aka IC Rating Scale, RCIF = Reference Class-Informed Forecasting, RGLM = Regression Using Generalized Linear Models, RST = Reasoning Stress Test, SPR = Standardizing Probability Ranges, STGC = Statistical Charts or Graphs, SWPC = SWARM Probability Calculator, TBLG = Table (Generic), TMLN = Timeline, VSIM = Visual Image (Other).

SUPPLEMENTARY MATERIALS B
Supplementary Materials for Chapter 3

Full ANCOVA Results from Main Text

We report the full results of the analyses of covariance (ANCOVA) for the two mixed design analyses on the dependent variable of min-max scaled, exponentially transformed Practical score reported in the main text, below.

Table SB1. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Block and Condition ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	640.08	2	320.04	0.83	.44	.00
AOMT	14387.46	1	14387.46	37.50	< .001	.06
Residuals	232485.42	606	383.64			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	737.39	5	147.48	1.22	.30	.00
Block × Condition	502.85	10	50.29	0.42	.94	.00
Block × AOMT	820.38	5	164.08	1.36	.24	.00
Residuals	365285.66	3030	120.56			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	737.39	4.92	149.88	1.22	.30	.00
Block × Condition	502.85	9.84	51.10	0.42	.94	.00
Block × AOMT	820.38	4.92	166.74	1.36	.24	.00
Residuals	365285.66	2981.55	122.52			

Note. Mauchly's test of sphericity was statistically significant ($p < .05$)

Table SB2. Experiment 2 Min-Max Scaled Exponentially Transformed Practical Score by Block and Condition ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	6258.30	2	3129.15	2.17	.12	.01
AOMT	26775.88	1	26775.88	18.54	< .001	.02
Residuals	1.252×10 ⁺⁶	867	1444.37			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	1614.87	4	403.72	1.38	.24	.00
Block × Condition	1872.57	8	234.07	0.80	.60	.00
Block × AOMT	1149.87	4	287.47	0.98	.42	.00
Residuals	1.016×10 ⁺⁶	3468	292.95			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	1614.87	3.92	412.11	1.38	.24	.00
Block × Condition	1872.57	7.84	238.94	0.80	.60	.00
Block × AOMT	1149.87	3.92	293.44	0.98	.42	.00
Residuals	1.016×10 ⁺⁶	3397.42	299.03			

Note. Mauchly's test of sphericity was statistically significant ($p < .05$)

Supplemental Analyses on Bias, Accuracy, and Confidence

In addition to the ANCOVAs on Practical Score reported in the main text, we also performed ANCOVAs on Bias, accuracy, and confidence across the six (or five) blocks of Experiment 1 (and 2). The ANCOVA results are presented below in Tables S3 – S4.

Table SB3. Experiment 1 Bias by Condition and Block ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	0.12	2	0.06	0.58	.56	.00
AOMT	4.67	1	4.67	44.98	< .001	.07
Residuals	62.87	606	0.10			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.13	5	0.03	0.92	.47	.00
Block × Condition	0.32	10	0.03	1.12	.35	.00
Block × AOMT	0.23	5	0.05	1.60	.16	.00
Residuals	86.08	3030	0.03			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.13	4.97	0.03	0.92	.47	.00
Block × Condition	0.32	9.93	0.03	1.12	.35	.00
Block × AOMT	0.23	4.97	0.05	1.60	.16	.00
Residuals	86.08	3009.70	0.03			

Table SB4. Experiment 2 Bias by Condition and Block ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	0.30	2	0.15	1.77	.17	.00
AOMT	2.22	1	2.22	26.00	< .001	.03
Residuals	74.14	867	0.09			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.05	4	0.01	0.88	.48	.00
Block × Condition	0.06	8	0.01	0.48	.87	.00
Block × AOMT	0.02	4	0.01	0.37	.83	4.295×10 ⁻⁴
Residuals	50.87	3468	0.02			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.05	3.95	0.01	0.88	.48	.00
Block × Condition	0.06	7.89	0.01	0.48	.87	.00
Block × AOMT	0.02	3.95	0.01	0.37	.83	4.295×10 ⁻⁴
Residuals	50.87	3421.12	0.02			

Note. Mauchly's test of sphericity was statistically significant ($p < .05$)

Table SB5. Experiment 1 Accuracy by Condition and Block ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	0.08	2	0.04	1.07	.35	.00
AOMT	1.54	1	1.54	42.07	< .001	.07
Residuals	22.15	606	0.04			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.11	5	0.02	0.80	.55	.00
Block $\times$ Condition	0.33	10	0.03	1.25	.26	.00
Block $\times$ AOMT	0.20	5	0.04	1.49	.19	.00
Residuals	80.18	3030	0.03			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.11	4.98	0.02	0.80	.55	.00
Block $\times$ Condition	0.33	9.96	0.03	1.25	.26	.00
Block $\times$ AOMT	0.20	4.98	0.04	1.49	.19	.00
Residuals	80.18	3016.80	0.03			

Table SB6. Experiment 2 Accuracy by Condition and Block ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	0.06	2	0.03	1.68	.19	.00
AOMT	1.55	1	1.55	80.80	< .001	.09
Residuals	16.63	867	0.02			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.03	4	0.01	0.56	.69	6.418×10^{-4}
Block $\times$ Condition	0.02	8	0.00	0.16	1.00	3.787×10^{-4}
Block $\times$ AOMT	0.03	4	0.01	0.49	.74	5.658×10^{-4}
Residuals	45.60	3468	0.01			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.03	3.97	0.01	0.56	.69	6.418×10^{-4}
Block $\times$ Condition	0.02	7.94	0.00	0.16	1.00	3.787×10^{-4}
Block $\times$ AOMT	0.03	3.97	0.01	0.49	.74	5.658×10^{-4}
Residuals	45.60	3442.19	0.01			

Table SB7. Experiment 1 Confidence by Condition and Block ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	0.20	2	0.10	1.43	0.24	0.01
AOMT	0.85	1	0.85	11.88	< .001	0.02
Residuals	43.20	606	0.07			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.00	5	8.891×10^{-4}	0.39	0.86	6.461×10^{-4}
Block $\times$ Condition	0.02	10	0.00	1.06	0.39	0.00
Block $\times$ AOMT	0.02	5	0.01	2.12	0.06	0.00
Residuals	6.88	3030	0.00			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.00	3.87	0.00	0.39	0.81	6.461×10^{-4}
Block $\times$ Condition	0.02	7.75	0.00	1.06	0.39	0.00
Block $\times$ AOMT	0.02	3.87	0.01	2.12	0.08	0.00
Residuals	6.88	2347.15	0.00			

Note. Mauchly's test of sphericity was statistically significant ($p < .05$)

Table SB8. Experiment 2 Confidence by Condition and Block ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	0.59	2	0.30	4.71	0.01	0.01
AOMT	0.06	1	0.06	0.97	0.33	0.00
Residuals	54.47	867	0.06			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.05	4	0.01	7.18	< .001	0.01
Block × Condition	0.02	8	0.00	1.57	0.13	0.00
Block × AOMT	0.01	4	0.00	1.12	0.35	0.00
Residuals	6.31	3468	0.00			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	0.05	2.54	0.02	7.18	< .001	0.01
Block × Condition	0.02	5.08	0.01	1.57	0.16	0.00
Block × AOMT	0.01	2.54	0.00	1.12	0.34	0.00
Residuals	6.31	2202.55	0.00			

Note. Mauchly's test of sphericity was statistically significant ($p < .05$)

Table SB9. Experiment 2 Confidence Post Hoc Comparisons for Training Condition

		Mean Difference	95% CI for Mean Difference		SE	<i>t</i>	Cohen's <i>d</i>	95% CI for Cohen's <i>d</i>		<i>p_{holm}</i>
Condition			Lower	Upper				Lower	Upper	
Control	Outcome	0.01	-0.02	0.03	0.01	0.68	0.05	-0.14	0.24	0.498
	Performance	-0.02	-0.04	0.00	0.01	2.23	-0.18	-0.37	0.01	0.052
Outcome	Performance	-0.03	-0.05	-0.01	0.01	2.94	-0.23	-0.42	-0.04	0.010

Note. *P*-value and confidence intervals adjusted for comparing a family of 3 estimates using the Holm method (confidence intervals corrected using the Bonferroni method). Results are averaged over the levels of: Block

Table SB10. Experiment 2 Confidence Post Hoc Comparisons for Block

		Mean Difference	95% CI for Mean Difference		SE	<i>t</i>	Cohen's <i>d</i>	95% CI for Cohen's <i>d</i>		<i>p_{holm}</i>
Block			Lower	Upper				Lower	Upper	
1	2	0.01	0.00	0.02	0.00	4.57	0.08	0.03	0.13	< .001
	3	0.01	0.01	0.02	0.00	5.59	0.10	0.05	0.15	< .001
	4	0.01	0.00	0.02	0.00	5.00	0.09	0.04	0.14	< .001
	5	0.02	0.01	0.02	0.00	8.22	0.14	0.09	0.19	< .001
2	3	0.00	0.00	0.01	0.00	1.02	0.02	-0.03	0.07	0.928
	4	8.783×10 ⁻⁴	-0.01	0.01	0.00	0.43	0.01	-0.04	0.06	1.000
	5	0.01	0.00	0.01	0.00	3.64	0.06	0.01	0.11	0.002
3	4	0.00	-0.01	0.01	0.00	-0.59	-0.01	-0.06	0.04	1.000
	5	0.01	-41.10	0.01	0.00	2.63	0.05	0.00	0.09	0.035
4	5	0.01	8.289×10 ⁻⁴	0.01	0.00	3.21	0.06	0.01	0.10	0.007

Note. *P*-value and confidence intervals adjusted for comparing a family of 10 estimates using the Holm method (confidence intervals corrected using the Bonferroni method). Results are averaged over the levels of: Condition

Table SB11. Experiment 2 Confidence Descriptive Statistics

Block	Condition	<i>N</i>	Mean	<i>SD</i>	<i>SE</i>	Coefficient of variation
1	Control	284	0.70	0.11	0.01	0.15
	Outcome	296	0.70	0.11	0.01	0.16
	Performance	291	0.73	0.11	0.01	0.15
2	Control	284	0.69	0.11	0.01	0.16
	Outcome	296	0.69	0.12	0.01	0.17
	Performance	291	0.71	0.12	0.01	0.17
3	Control	284	0.69	0.12	0.01	0.17
	Outcome	296	0.69	0.12	0.01	0.18
	Performance	291	0.71	0.12	0.01	0.17
4	Control	284	0.69	0.12	0.01	0.17
	Outcome	296	0.69	0.13	0.01	0.19
	Performance	291	0.71	0.12	0.01	0.17
5	Control	284	0.69	0.12	0.01	0.17
	Outcome	296	0.68	0.12	0.01	0.18
	Performance	291	0.70	0.12	0.01	0.18

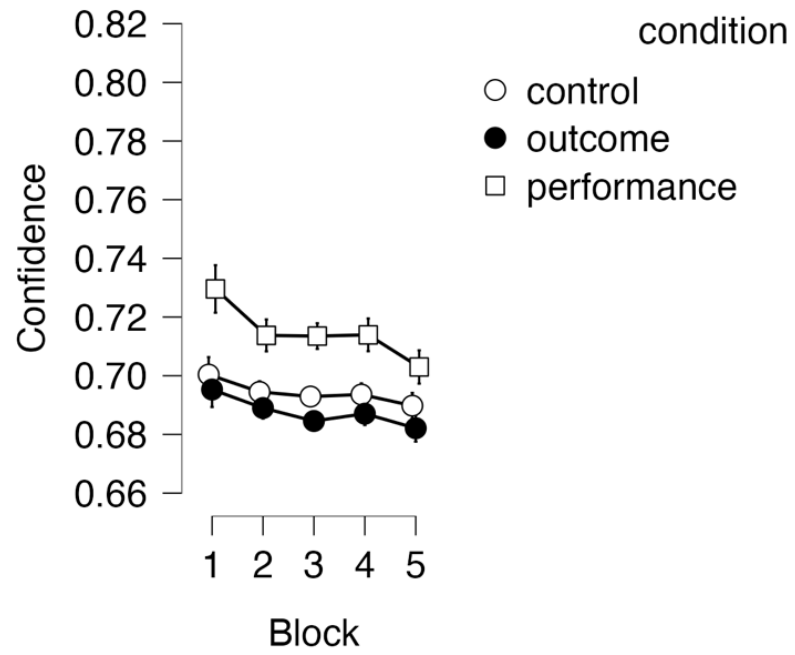


Figure SB1. Experiment 2 Confidence by Block

Test Block Analyses

In addition to the ANCOVAs on Practical Score across all blocks of the experiments reported in the main text, we also performed ANCOVAs on transformed Practical score on only the test blocks (for Experiment 1, we took the average score of the two test blocks). The ANCOVA results are presented below in Tables SB12-SB13, below.

Table SB12. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Condition ANCOVA Results

	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
AOMT	2443.26	1	2443.26	23.78	< .001	0.04
Condition	312.19	2	156.09	1.52	0.22	0.01
Residuals	62261.35	606	102.74			

Table SB 13. Experiment 2 Min-Max Scaled Exponentially Transformed Practical Score by Condition ANCOVA Results

	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
AOMT	5818.49	1	5818.49	11.85	< .001	0.01
Condition	613.79	2	306.90	0.63	0.54	0.00
Residuals	425723.91	867	491.03			

Sex-Based Analyses

Table SB14. Experiment 1 Min-Max Scaled Exponentially Transformed Practical Score by Condition, Block, and Sex ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	641.90	2	320.95	0.85	0.43	0.00
AOMT	13497.05	1	13497.05	35.88	< .001	0.06
Sex	4512.60	1	4512.60	12.00	< .001	0.02
Condition x Sex	1090.07	2	545.04	1.45	0.24	0.01
Residuals	226837.67	603	376.18			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	763.76	5	152.75	1.27	0.27	0.00
Block x Condition	511.49	10	51.15	0.43	0.94	0.00
Block x AOMT	849.87	5	169.97	1.41	0.22	0.00
Block x Sex	441.07	5	88.21	0.73	0.60	0.00
Block x Condition x Sex	1903.44	10	190.34	1.58	0.11	0.01
Residuals	362948.25	3015	120.38			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	763.76	4.92	155.14	1.27	0.28	0.00
Block x Condition	511.49	9.85	51.95	0.43	0.93	0.00
Block x AOMT	849.87	4.92	172.63	1.41	0.22	0.00
Block x Sex	441.07	4.92	89.59	0.73	0.60	0.00
Block x Condition x Sex	1903.44	9.85	193.32	1.58	0.11	0.01
Residuals	362948.25	2968.58	122.26			

Note. Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Table SB15. Experiment 2 Min-Max Scaled Exponentially Transformed Practical Score by Condition, Block, and Sex ANCOVA Results

Between Subjects Effects						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Condition	7164.12	2	3582.06	2.54	0.08	0.01
AOMT	32550.15	1	32550.15	23.11	< .001	0.03
Sex	34993.76	1	34993.76	24.84	< .001	0.03
Condition x Sex	2333.74	2	1166.87	0.83	0.44	0.00
Residuals	1.217×10 ⁺⁶	864	1408.76			
Within Subjects Effects (No Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	2061.46	4	515.36	1.76	0.13	0.00
Block x Condition	2158.39	8	269.80	0.92	0.50	0.00
Block x AOMT	2724.66	4	681.16	2.33	0.05	0.00
Block x Sex	1211.71	4	302.93	1.04	0.39	0.00
Block x Condition x Sex	2063.70	8	257.96	0.88	0.53	0.00
Residuals	1.011×10 ⁺⁶	3456	292.58			
Within Subjects Effects (Greenhouse-Geisser Correction)						
	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2_p
Block	2061.46	3.92	526.19	1.76	0.14	0.00
Block x Condition	2158.39	7.84	275.47	0.92	0.50	0.00
Block x AOMT	2724.66	3.92	695.48	2.33	0.06	0.00
Block x Sex	1211.71	3.92	309.29	1.04	0.39	0.00
Block x Condition x Sex	2063.70	7.84	263.38	0.88	0.53	0.00
Residuals	1.011×10 ⁺⁶	3384.87	298.73			

Note. Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Control Checks and Preliminary Analyses

Urban Agglomeration List

We report the results of the control checks involving urban agglomeration (UA) list on the dependent variables of transformed Practical score and Bias for Experiments 1 and 2, followed by the relevant control checks on actively open-minded thinking (AOMT) and the engagement question in Experiment 2.

To test whether city list affected the results of the primary analyses, mixed design ANOVAs were performed with UA list as a between-subjects variable and the six (or five) blocks of Experiment 1 (or 2) as a within-subjects variable. A main effect of UA list might indicate that the choice of UAs impacted the overall difficulty of the test, and an interaction effect might indicate that this difference impacted learning calibration. Recall that the Practical score of each block was min-max scaled and exponentially transformed with base 100 or 200 for Experiments 1 and 2, respectively.

For Experiment 1, the Greenhouse-Geisser corrected ($\varepsilon = .98$) ANOVA revealed no statistically significant main effect of UA list on transformed Practical score, $F(1, 608) = 1.49, p = 0.22, \eta^2_p = .00$, nor was there a statistically significant UA list x block interaction, $F(4.92, 2989.15) = 1.08, p = .37, \eta^2_p = .00$. For the Bias ANOVA, there was no statistically significant main effect of UA list on Bias, $F(1, 608) = 0.02, p = .89, \eta^2_p = .00$, nor was there a significant UA list x block order interaction, $F(5, 3040) = 1.16, p = .33, \eta^2_p = .00$.

For Experiment 2, the Greenhouse-Geisser corrected ($\varepsilon = .98$) ANOVA on Practical score revealed no statistically significant main effect of UA list, $F(1, 869) = 0.03, p = .86, \eta^2_p = .00$, nor was there a statistically significant UA list x block order interaction, $F(3.92, 3407.04) = 1.49, p = 0.20, \eta^2_p = .00$. The Greenhouse-Geisser corrected ($\varepsilon = .99$) ANOVA with Bias as the dependent variable revealed no statistically significant main effect of UA list, $F(1, 869) = 0.06, p = .80, \eta^2_p = .00$ nor was there a statistically significant interaction effect between UA list and block order, $F(3.95, 3432.84) = 0.56, p = .69, \eta^2_p = .00$. Based on these results, the two lists did not appear to affect the main dependent variables of interest, so they were combined in all subsequent analyses.

AOMT

To ensure that AOMT did not differ across the three training conditions of both experiments, we conducted between-subjects ANOVAs with AOMT as the dependent variable. For Experiment 1, the analysis showed no statistically significant difference in AOMT between the three training groups, $F(2, 607) = 1.29, p = .28, \eta^2_p = .00$. Likewise for Experiment 2, the analysis revealed no statistically significant differences in AOMT between the three training groups, $F(2, 868) = 0.79, p = .45, \eta^2_p = .00$. The results suggest that AOMT was not differentially distributed across the treatment conditions and was thus appropriate for inclusion as a covariate in both experiments.

Experiment 2 Engagement Question

The characteristics of Participants who responded 'No' to the engagement question in Experiment 2 are examined relative to those who answered 'Yes' (referred to as engagement group) here. First, to examine whether there was a systematic relationship between engagement group and training condition, a χ^2 test of independence was conducted. The test revealed no statistically significant relationship between the two variables, $\chi^2(2, N = 908) = 1.20, p = .55$, Cramer's $V = .04$, indicating that training condition did not significantly influence participants' answer to the engagement question.

To determine whether engagement group affected the primary analyses, we compared the two groups across the five blocks of the experiment using mixed design ANOVAs with transformed Practical score and Bias as the dependent variables. The Greenhouse-Geisser corrected ($\varepsilon = .98$) ANOVA on transformed Practical score revealed no statistically significant main effect of engagement group, $F(1, 906) = 0.40, p = .53, \eta^2_p = .00$, nor was there a statistically significant engagement group x block interaction, $F(3.91, 3544.59) = 1.74, p = .14, \eta^2_p = .00$. The Greenhouse-Geisser corrected ($\varepsilon = .98$) ANOVA on Bias found no statistically significant effect of engagement group, $F(1, 906) = .26, p = .61, \eta^2_p = .00$, nor was there a statistically significant engagement group x block interaction effect, $F(3.93, 3562.58) = 0.37, p = .82, \eta^2_p = .00$. The results suggest no substantial differences between the two engagement groups. Their data were excluded from all subsequent analyses nonetheless since they admitted to not paying attention during parts of the experiment.

Outlier Analyses

Outlier analyses were conducted on the Practical score variable due to its negative skew caused by a long tail of extremely low scores. The outlier analyses for the average of all blocks of Experiment 1 are presented first, followed by those for Experiment 2. We also include analyses conducted on the average untransformed Practical score of the test blocks in Experiments 1 and 2 (called test Practical score), as they shed some light on participant engagement at the end of each experiment, even though test block analyses were not reported in the main text due to redundancy with the analyses including all blocks.

Experiment 1 – Practical Score Across Blocks

The Practical score variable across the six blocks of Experiment 1 showed considerable negative skew (range -1.63 - -3.34) and leptokurtosis (range: 3.66 – 19.30). Removal of extreme scores where the average score $z \leq -1.645$ was not desirable, as extreme scores were significantly above chance performance ($M = .55$, $SD = .07$), one-tailed $t(51) = 5.89$, $p < .001$, $d = 0.82$, 95% CI [.54, .57]. Even participants where $z \leq -3.09$ were above chance accuracy ($M = .56$, $SD = .05$), one-tailed $t(13) = 3.88$, $p < .001$, $d = 1.04$, two-tailed 95% CI [.53, .58], so their removal may have implied removing participants who were compliant with the experiment.

To address the issue of negative skew and leptokurtosis, the data were exponentially transformed with base 100 after first being min-max scaled. The resulting data showed much less skew (range: -0.63 – 0.52) and leptokurtosis (range: -.01 - 3.61) than the raw scores. The full ANCOVA results after transformation are reported in Table S1.

Experiment 2 – Practical Score Across Blocks

As was the case in Experiment 1, the Practical score data in Experiment 2 were highly negatively skewed (skew range: -2.13 - -3.14) and extremely highly leptokurtotic (kurtosis range: 4.53 - 13.75) due to a long tail of extremely low scores. An outlier analysis of participants with the most extremely low average Practical score ($z \leq -3.09$) revealed that they had marginally statistically significantly higher than chance accuracy ($M = 0.52$, $SD = 0.05$), $t(16) = 1.75$, one tailed $p = .050$, $d = 0.42$, two tailed 95% CI [-.004, .044]. Given that these participants were possibly giving valid data, we decided to transform the raw Practical score values in each block using the exponential function with base 200, after the scores had first been

min-max scaled. The resulting transformed values were less negatively skewed (skew range: -0.32 - -0.90), and less leptokurtotic (kurtosis range: 0.19 – 1.97) than the original scores. The results of ANCOVA performed on the transformed data are reported in Table S2.

Experiment 1 – Test Practical Score

We also examined average Practical score across the two test blocks of Experiment 1 (i.e., test Practical score) to investigate whether the effect of training would transfer to a condition without feedback. An inspection of the distributions of test Practical score across the three treatment groups showed that it was highly negatively skewed (skew range: -3.08 - -1.86) and leptokurtic (kurtosis range: 3.19 – 13.78) due to a long tail of extremely low Practical scores. Participants with $z \leq -3.09$ of final Practical score variable within each treatment group were examined to determine if they were noncompliant with the instructions and thus possible candidates for exclusion.

The outlier analysis suggested that the extreme scorers ($n = 15$) had average accuracy scores in the final two blocks that were not statistically significantly above chance, ($M = .48$, $SD = .08$), $t(14) = -1.07$, one-tailed $p = .85$, $d = -0.28$, two-tailed 95% CI [-0.07, .02]. By contrast, non-extreme scorers ($n = 595$) were significantly above chance accuracy with a large effect size ($M = 0.60$, $SD = 0.12$), one-tailed $t(594) = 20.20$, $p < .001$, $d = 0.83$, two-tailed 95% CI [0.10, 0.13].

Despite their low accuracy, extreme scorers had an extremely high average confidence ($M = .97$, $SD = 0.04$) compared non-extreme scorers ($M = .71$, $SD = .11$). Due a violation of the assumptions of normality and inequality of variances, and due to the small and unequally sized groups, a Mann-Whitney U test was conducted on the difference in confidence between these two groups. The test found that extreme scorers had a statistically significantly higher mean confidence compared to non-extreme scorers, $W(608) = 166.00$, $p < .001$, $r_{\text{rank-biserial}} = -.96$, 95% CI [-0.32, -0.22]. Finally, being an extreme scorer did not appear to be systematically related to the training condition a participant was assigned to, as a χ^2 test of independence was not statistically significant, $\chi^2(2, N = 610) = 2.77$, $p = .25$, Cramer's $V = .07$. Thus, there was some evidence to suggest that extreme scorers were not compliant with the task, as they had extremely high confidence, accuracy marginally above chance, and were not systematically related to the training condition. After removal of the extreme scorers, the skewness (skew range: -1.51 - -1.64) and leptokurtosis (kurtosis range: 2.07 – 3.29) of test Practical score were closer to normal ranges.

Experiment 2 – Test Practical Score

Outliers on Practical score in the test block of Experiment 2 were also analyzed as above. Test Practical score in Experiment 2 was negatively skewed (skewness range: -2.88 - -2.31) with an extreme amount of leptokurtosis (kurtosis range: 6.86 – 10.14), largely owing to a long tail of low-scoring outliers. An outlier analysis of participants with test Practical score $z \leq -3.09$ within a condition revealed that their accuracy was not statistically higher than chance ($M = .50$, $SD = 0.09$), $t(10) = -0.18$, one-tailed $p = .57$, $d = -0.05$, two-tailed 95% CI [-0.06, 0.05]. Non-extreme scorers had accuracy that was statistically significantly higher than chance ($M = .59$, $SD = 0.12$), $t(859) = 21.62$, one-tailed $p < .001$, $d = .74$, two-tailed 95% CI [.08, .10]. Despite their low accuracy, the extreme scorers had an extremely high confidence ($M = .96$, $SD = .05$) that was statistically significantly higher than non-extreme participants ($M = .69$, $SD = 0.12$), Mann-Whitney $U = 212.50$, $p < .001$, $r_{\text{rank-biserial}} = -.96$, 95% CI [-0.35, -0.22]. Furthermore, being an extreme scorer was not systematically related to belonging to one of the training conditions, $\chi^2(2, N = 871) = 0.15$, $p = .93$, Cramer's $V = 0.01$. After removal of extreme scorers, test Practical score was less skewed (skew range: -2.31- -1.64) and leptokurtotic (kurtosis range: 2.60 – 6.32).

Bayesian Analyses

The following JASP settings for model specifications were used: the prior on coefficients was set for fixed and random terms with r scale fixed effects = 0.50, r scale random effects = 1.00, and r scale covariates = 0.354. Numerical accuracy was set manually to 10,000 samples and nuisance was hidden in the model. The integration method was set to automatic. The number of posterior samples was set manually to 1000. The principle of marginality was enforced for fixed effects but not for random slopes. The model prior was set to uniform (equiprobable). The outputs of the JASP analyses for min-max scaled exponentially transformed Practical Score and Bias for Experiments 1 and 2 are presented in Tables S5-S8, below.

Table SB16. Experiment 1 Bayesian ANOVA on Min-Max Scaled Exponentially Transformed Practical Score by Condition and Block

Models	P(M)	P(M data)	BF _M	BF ₀₁	Error %
Null model (incl. subject and random slopes)	0.20	0.97	113.31	1.00	
Condition	0.20	0.03	0.14	28.38	0.84
Block	0.20	5.635×10^{-5}	2.254×10^{-4}	17141.39	1.66
Block + Condition	0.20	1.960×10^{-6}	7.839×10^{-6}	492883.79	1.19
Block + Condition + Block x Condition	0.20	1.427×10^{-11}	5.710×10^{-11}	$6.767 \times 10^{+10}$	1.33

Note. All models include subject, and random slopes for all repeated measures factors.

Table SB17. Experiment 1 Bayesian ANOVA on Bias by Condition and Block

Models	P(M)	P(M data)	BF _M	BF ₀₁	Error %
Null model (incl. subject and random slopes)	0.20	0.97	113.97	1.00	
Condition	0.20	0.03	0.14	28.52	1.32
Block	0.20	2.759×10^{-5}	1.104×10^{-4}	35017.00	0.47
Block + Condition	0.20	9.888×10^{-7}	3.955×10^{-6}	976987.91	2.62
Block + Condition + Block $\times$ Condition	0.20	1.505×10^{-10}	6.022×10^{-10}	$6.417 \times 10^{+9}$	1.17

Note. All models include subject, and random slopes for all repeated measures factors.

Table SB18. Experiment 2 Bayesian ANOVA on Min-Max Scaled Exponentially Transformed Practical Score by Condition and Block

Models	P(M)	P(M data)	BF _M	BF ₀₁	Error %
Null model (incl. subject and random slopes)	0.20	0.90	36.68	1.00	
Condition	0.20	0.08	0.35	11.20	0.51
Block	0.20	0.02	0.07	55.08	0.91
Block + Condition	0.20	0.00	0.01	621.38	0.82
Block + Condition + Block $\times$ Condition	0.20	6.457×10^{-8}	2.583×10^{-7}	$1.397 \times 10^{+7}$	4.00

Note. All models include subject, and random slopes for all repeated measures factors.

Table SB19. Experiment 2 Bayesian ANOVA on Bias by Condition and Block

Models	P(M)	P(M data)	BF _M	BF ₀₁	Error %
Null model (incl. subject and random slopes)	0.20	0.93	52.01	1.00	
Condition	0.20	0.07	0.31	13.06	0.53
Block	0.20	2.848×10^{-4}	0.00	3260.88	0.43
Block + Condition	0.20	2.227×10^{-5}	8.908×10^{-5}	41696.43	2.67
Block + Condition + Block $\times$ Condition	0.20	3.007×10^{-10}	1.203×10^{-9}	$3.088 \times 10^{+9}$	0.94

Note. All models include subject, and random slopes for all repeated measures factors.

Protocol for Experiment 1

Protocol for Calibration Training Experiment

[Participants first complete a series of target audience questions to determine if they qualify for the study]

Are you -

[Male / Female / Prefer not to say]

What is your age? Please select the age-bracket that contains your current age.

[0-17 / 18+]

What citizenship(s) do you hold?

[American / Canadian / European / Mexican / Other]

What is your first language?

[Cantonese / English / French / Mandarin / Spanish / Other]

[Participants then fill out the informed consent form]

[Participants are then given the following introductory instructions]

Thank you for participating in this study!

We care about the quality of our survey data and hope to receive the most accurate measures of your knowledge, so it is important to us that you thoughtfully provide your best answer to each question in the survey.

Before beginning, please make sure to minimize any distractions so you can give the study your full attention. This study will take approximately 25 minutes to complete.

Do you commit to minimizing distractions and providing your thoughtful and honest answers to the questions in this survey?

- I will minimize distractions and provide my best answers
- I can't commit to 25 minutes of undistracted time
- I can't promise either way

[Page break]

For this study, you will be answering questions that ask you to decide which of two cities has a higher population.

You will also be asked to indicate how confident you are that each of your answers are correct on a scale from 50% to 100%.

Please note that 50% would indicate a pure guess, and 100% would indicate absolute certainty.

Questions will be grouped in "blocks" of 10, and there are 6 blocks in total.

Afterward, we will ask you about your level of knowledge of each city, followed by some questions about your thinking style.

Please answer to the best of your ability and provide honest indications of your confidence.

Please do not look up the answers, e.g., on the internet.

- I will not look up any of the answers
- I might look up some of the answers

[Outcome feedback condition instructions]

After you answer each question, you will receive feedback on whether or not you were correct.

Additionally, as noted on the previous screen, questions will be "blocked" into sets of 10. After each of the first four blocks, you will receive feedback on your overall accuracy rate for each completed block (namely, for each of the four sets of 10 questions). For example, at the end of Block 3, you will see something like this:

Block number	% Correct
1	60
2	70
3	50

In the final two blocks, you will receive no feedback at all.

[Performance feedback condition instructions]

After you answer each question, you will receive feedback on whether or not you were correct. You will also receive a score based on your degree of correctness and confidence. You get the best possible score

when you are maximally confident and correct, and you get the worst possible score when you are maximally confident and incorrect.

Additionally, as noted on the previous screen, questions will be "blocked" into sets of 10. After each of the first four blocks, you will receive feedback on your overall accuracy rate for each completed block (namely, for each of the four sets of 10 questions). Your total score for that block will be displayed next to your accuracy rate. You will also receive a "bias" score indicating whether you were generally over-confident, under-confident, or well-calibrated (namely, unbiased) for each completed block. For example, at the end of Block 3, you will see something like this:

Block number	% Correct	Score (max 100)	Bias	Bias Interpretation
1	60	20.55	0.3	over-confident
2	70	25.65	-0.2	under-confident
3	50	15.85	0.004	well-calibrated

A positive bias score indicates over-confidence, namely, that your average confidence is more than your average accuracy in that block.

A negative bias score indicates under-confidence, namely, that your average confidence is less than your average accuracy in that block.

A bias of 0 indicates that you were perfectly well-calibrated, namely, that your average confidence is matched to your average accuracy in that block. For instance, if you were 70% confident on average and got 7 out of 10 problems correct, you would be perfectly calibrated.

You should try to get your bias score as close to zero as possible on each block, as this will also increase your overall score for the block. If you were over-confident, you should generally reduce how confident you say you are in the next block. If you were under-confident, you should generally increase how confident you are in the next block.

In the final two blocks, you will receive no feedback at all.

[Stimuli for all conditions]

[City A, Country A] had a larger population than [City B, Country B] in 2020.
True / False

How confident are you that your answer is correct?
[50 – 100% slider]

[Feedback for correct answer in Outcome Feedback condition trial]
Correct!

[Feedback for incorrect answer in Outcome Feedback condition trial]
Incorrect!

[Feedback for correct answer in Performance Feedback condition trial]

Correct! You were [X]% confident.
+ [X] points!

[Feedback for incorrect answer in Performance Feedback condition trial]
Incorrect! You were [X]% confident.
- [X] points!

[If participants are in the Outcome Feedback or Performance Feedback conditions, after each block of 10 questions, participants are presented with a summary of their performance (as depicted in the instructions above). If they are in the control condition, they are simply informed that they have completed a block]

[After completing 4 blocks, participants complete the Sports Participation Question, which serves as an attention-check item]:

Sports Activities You Engage in Regularly

Most modern theories of decision making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. In order to facilitate our research on decision making we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So, in order to demonstrate that you have read the instructions, please ignore the sports items below. Instead, simply click on the red button to proceed to the next screen. Thank you very much.

Which of these activities do you engage in regularly? (click on all that apply)

- | | |
|------------------------------------|----------------------------------|
| ☐ skiing | ☐ Football |
| ☐ soccer | ☐ Swimming |
| ☐ snowboarding | ☐ Tennis |
| ☐ running | ☐ Basketball |
| ☐ hockey | ☐ Cycling |

[After completing 4 training blocks, if participants are in the Outcome Feedback or Performance Feedback conditions, participants will receive the following message]

In this part of the study, you will once again be answering questions about city populations and providing confidence ratings, but there will no longer be any feedback.

[Participants will then complete two more blocks without any feedback]

[Following the main judgement task, participants will indicate their recognition of and familiarity with each of the (60+) cities they were presented with. They will receive the following instructions]

Thank you for your answers!

For this part of the study, you will be asked to indicate your level of knowledge about particular cities prior to this study.

Please answer as honestly as possible.

[The recognition question is like the below]

Please indicate your level of knowledge about the city of **[City A, Country A]**.

[I have never heard of it prior to this study / I have only heard the name before / I know a little about the city / I am moderately knowledgeable about the city / I am very knowledgeable about the city]

[Next, participants complete the Actively Open Minded Thinking questions, in random order]

Please respond to the following questions by using the following 5-point scale:

–2 = Strongly disagree

–1 = Slightly disagree

0 = Agree/Disagree about equally

1 = Slightly agree

2 = Strongly agree

1. Allowing oneself to be convinced by an opposing argument is a sign of good character.
2. People should take into consideration evidence that goes against their beliefs.
3. People should revise their beliefs in response to new information or evidence.
4. Changing your mind is a sign of weakness.
5. Intuition is the best guide in making decisions.
6. It is important to persevere in your beliefs even when evidence is brought to bear against them.
7. One should disregard evidence that conflicts with one's established beliefs.
8. People should search actively for reasons why their beliefs might be wrong.

[Next, participants complete the following demographics questions]

Are you –

- Male
- Female
- Prefer not to say

What is the highest level of education that you have completed?

- High School (or equivalent)

- Trade School (non-military)
- University - Some undergraduate
- University – Undergraduate
- University - Graduate (Masters)
- University – PhD

How old are you (in years)?

[0 – 100 slider response]

Please take a moment to share any feedback you may have about this survey. Remember - your responses will remain anonymous.

[text response]

[Finally, participants are debriefed]

Protocol for Experiment 2

[Participants first complete a series of target audience questions to determine if they qualify for the study]

Are you -

[Male / Female / Prefer not to say]

What is your age? Please select the age-bracket that contains your current age.

[0-17 / 18+]

What citizenship(s) do you hold?

[American / Canadian / European / Mexican / Other]

What is your first language?

[Cantonese / English / French / Mandarin / Spanish / Other]

[Participants then fill out the informed consent form]

[Participants are then given the following introductory instructions]

Thank you for participating in this study!

We care about the quality of our survey data and hope to receive the most accurate measures of your knowledge, so it is important to us that you thoughtfully provide your best answer to each question in the survey.

Before beginning, please make sure to minimize any distractions so you can give the study your full attention. This study will take approximately 25 minutes to complete.

Do you commit to minimizing distractions and providing your thoughtful and honest answers to the questions in this survey?

- I will minimize distractions and provide my best answers
- I can't commit to 25 minutes of undistracted time
- I can't promise either way

[Page break]

For this study, you will be answering questions that ask you to decide which of two cities has a higher population.

You will also be asked to indicate how confident you are that each of your answers are correct on a scale from 50% to 100%.

Please note that 50% would indicate a pure guess, and 100% would indicate absolute certainty.

Questions will be grouped in "blocks" of 20, and there are 5 blocks in total.

Afterward, we will ask you some questions about your thinking style.

Please answer to the best of your ability and provide honest indications of your confidence.

Please do not look up the answers, e.g., on the internet.

- I will not look up any of the answers
- I might look up some of the answers

[Outcome feedback condition instructions]

After you answer each question, you will receive feedback on whether or not you were correct.

Additionally, as noted on the previous screen, questions will be "blocked" into sets of 20. After each of the first four blocks, you will receive feedback on your overall accuracy rate for each completed block (namely, for each of the four sets of 20 questions). For example, at the end of Block 3, you will see something like this:

Block number	% Correct
1	60
2	70
3	50

In the final block, you will receive no feedback at all.

[Performance feedback condition instructions]

After you answer each question, you will receive feedback on whether or not you were correct. You will also receive a score based on your degree of correctness and confidence. You get the best possible score when you are maximally confident and correct, and you get the worst possible score when you are maximally confident and incorrect.

Additionally, as noted on the previous screen, questions will be "blocked" into sets of 20. After each of the first four blocks, you will receive feedback on your overall accuracy rate for each completed block (namely, for each of the four sets of 20 questions). Your total score for that block will be displayed next to your accuracy rate. You will also receive a "bias" score indicating whether you were generally over-confident, under-confident, or well-calibrated (namely, unbiased) for each completed block. For example, at the end of Block 3, you will see something like this:

Block number	% Correct	Score (max 100)	Bias	Bias Interpretation
1	60	20.55	0.3	over-confident
2	70	25.65	-0.2	under-confident
3	50	15.85	0.004	well-calibrated

A positive bias score indicates over-confidence, namely, that your average confidence is more than your average accuracy in that block.

A negative bias score indicates under-confidence, namely, that your average confidence is less than your average accuracy in that block.

A bias of 0 indicates that you were perfectly well-calibrated, namely, that your average confidence is matched to your average accuracy in that block. For instance, if you were 70% confident on average and got 14 out of 20 problems correct, you would be perfectly calibrated.

You should try to get your bias score as close to zero as possible on each block, as this will also increase your overall score for the block. If you were over-confident, you should generally reduce how confident you say you are in the next block. If you were under-confident, you should generally increase how confident you are in the next block.

In the final two blocks, you will receive no feedback at all.

[Prior to completing the training blocks, participants complete the Sports Participation Question, which serves as an attention-check item]:

Sports Activities You Engage in Regularly

Most modern theories of decision making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. In order to facilitate our research on decision making we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So, in order to demonstrate that you have read the instructions, please ignore the sports items below. Instead, simply click on the red button to proceed to the next screen. Thank you very much.

Which of these activities do you engage in regularly? (click on all that apply)

- | | |
|------------------------------------|----------------------------------|
| ☐ skiing | ☐ Football |
| ☐ soccer | ☐ Swimming |
| ☐ snowboarding | ☐ Tennis |
| ☐ running | ☐ Basketball |
| ☐ hockey | ☐ Cycling |

[Stimuli for all conditions]

[City A, Country A] had a larger population than [City B, Country B] in 2020.
True / False

How confident are you that your answer is correct?
[50 – 100% slider]

[Feedback for correct answer in Outcome Feedback condition trial]

Correct!

[Feedback for incorrect answer in Outcome Feedback condition trial]

Incorrect!

[Feedback for correct answer in Performance Feedback condition trial]

Correct! You were [X]% confident.

+ [X] points!

[Feedback for incorrect answer in Performance Feedback condition trial]

Incorrect! You were [X]% confident.

- [X] points!

[If participants are in the Outcome Feedback or Performance Feedback conditions, after each block of 20 questions, participants are presented with a summary of their performance (as depicted in the instructions above). If they are in the control condition, they are simply informed that they have completed a block]

[After completing 4 training blocks, if participants are in the Outcome Feedback or Performance Feedback conditions, participants will receive the following message]

In this part of the study, you will once again be answering questions about city populations and providing confidence ratings, but there will no longer be any feedback.

[Participants will then complete one more block without any feedback]

[Next, participants answer a question regarding their engagement in the study]

Did you find the task engaging enough so that you were able to pay attention for the entire time?
(You will not be penalized based on your answer).

- Yes
- No

[Next, participants complete the Actively Open Minded Thinking questions, in random order]

Please respond to the following questions by using the following 5-point scale:

–2 = Strongly disagree

–1 = Slightly disagree

0 = Agree/Disagree about equally

1 = Slightly agree

2 = Strongly agree

1. Allowing oneself to be convinced by an opposing argument is a sign of good character.
2. People should take into consideration evidence that goes against their beliefs.
3. People should revise their beliefs in response to new information or evidence.
4. Changing your mind is a sign of weakness.

5. Intuition is the best guide in making decisions.
6. It is important to persevere in your beliefs even when evidence is brought to bear against them.
7. One should disregard evidence that conflicts with one's established beliefs.
8. People should search actively for reasons why their beliefs might be wrong.

[Next, participants complete the following demographics questions]

Are you –

- Male
- Female
- Prefer not to say

What is the highest level of education that you have completed?

- High School (or equivalent)
- Trade School (non-military)
- University - Some undergraduate
- University – Undergraduate
- University - Graduate (Masters)
- University – PhD

How old are you (in years)?

[0 – 100 slider response]

Please take a moment to share any feedback you may have about this survey. Remember - your responses will remain anonymous.

[text response]

[Finally, participants are debriefed]

Table SB20. Names of Cities and Countries Used in Experiments 1 and 2

Rank	City	Country	Population (thousands)
1	Tokyo	Japan	37 393
2	Delhi	India	30 291
3	Shanghai	China	27 058
4	São Paulo	Brazil	22 043
5	Mexico City	Mexico	21 782
6	Dhaka	Bangladesh	21 006
7	Cairo	Egypt	20 901
8	Beijing	China	20 463
9	Mumbai	India	20 411
10	Osaka	Japan	19 165
11	New York-Newark	United States of America	18 804
12	Karachi	Pakistan	16 094
13	Chongqing	China	15 872
14	Istanbul	Turkey	15 190
15	Buenos Aires	Argentina	15 154
16	Kolkata	India	14 850
17	Lagos	Nigeria	14 368
18	Kinshasa	Democratic Republic of the Congo	14 342
19	Manila	Philippines	13 923
20	Tianjin	China	13 589
21	Rio de Janeiro	Brazil	13 458
22	Guangzhou, Guangdong	China	13 302
23	Lahore	Pakistan	12 642
24	Moscow	Russian Federation	12 538
25	Los Angeles-Long Beach-Santa Ana	United States of America	12 447
26	Shenzhen	China	12 357
27	Bangalore	India	12 327
28	Paris	France	11 017
29	Bogotá	Colombia	10 978
30	Chennai	India	10 971
31	Jakarta	Indonesia	10 770
32	Lima	Peru	10 719
33	Bangkok	Thailand	10 539
34	Hyderabad	India	10 004
35	Seoul	Republic of Korea	9 963
36	Nagoya	Japan	9 552
37	London	United Kingdom	9 304
38	Chengdu	China	9 136
39	Tehran	Iran (Islamic Republic of)	9 135
40	Chicago	United States of America	8 865
41	Nanjing, Jiangsu	China	8 847
42	Ho Chi Minh City	Viet Nam	8 602
43	Wuhan	China	8 365
44	Luanda	Angola	8 330

45	Ahmadabad	India	8 059
46	Xi'an, Shaanxi	China	8 001
47	Kuala Lumpur	Malaysia	7 997
48	Hangzhou	China	7 642
49	Hong Kong	China, Hong Kong SAR	7 548
50	Dongguan	China	7 408
51	Foshan	China	7 327
52	Riyadh	Saudi Arabia	7 231
53	Shenyang	China	7 220
54	Surat	India	7 185
55	Baghdad	Iraq	7 144
56	Suzhou, Jiangsu	China	7 070
57	Santiago	Chile	6 767
58	Dar es Salaam	United Republic of Tanzania	6 702
59	Pune	India	6 629
60	Madrid	Spain	6 618
61	Haerbin	China	6 387
62	Houston	United States of America	6 371
63	Dallas-Fort Worth	United States of America	6 301
64	Toronto	Canada	6 197
65	Miami	United States of America	6 122
66	Belo Horizonte	Brazil	6 084
67	Singapore	Singapore	5 935
68	Khartoum	Sudan	5 829
69	Atlanta	United States of America	5 803
70	Johannesburg	South Africa	5 783
71	Philadelphia	United States of America	5 717
72	Qingdao	China	5 620
73	Dalian	China	5 618
74	Barcelona	Spain	5 586
75	Kitakyushu-Fukuoka M.M.A.	Japan	5 529
76	Saint Petersburg	Russian Federation	5 468
77	Ji'nan, Shandong	China	5 360
78	Yangon	Myanmar	5 332
79	Zhengzhou	China	5 323
80	Washington, D.C.	United States of America	5 322
81	Alexandria	Egypt	5 281
82	Abidjan	Côte d'Ivoire	5 203
83	Guadalajara	Mexico	5 179
84	Ankara	Turkey	5 118
85	Chittagong	Bangladesh	5 020
86	Melbourne	Australia	4 968
87	Sydney	Australia	4 926
88	Monterrey	Mexico	4 874
89	Addis Ababa	Ethiopia	4 794
90	Nairobi	Kenya	4 735
91	Hà Noi	Viet Nam	4 678
92	Brasília	Brazil	4 646

93	Cape Town	South Africa	4 618
94	Jiddah	Saudi Arabia	4 610
95	Changsha	China	4 578
96	Phoenix-Mesa	United States of America	4 511
97	Kunming	China	4 443
98	Changchun	China	4 426
99	Xinbei	Taiwan	4 398
100	Ürümqi	China	4 369

REFERENCES

- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4(63).
<https://doi.org/10.12688/wellcomeopenres.15191.2>
- Andersen, D. B., Sunde, N., & Porter, K. (2024). *Tool Induced Biases? Misleading Data Presentation as a Biasing Source in Digital Forensic Analysis*. Social Science Research Network.
<https://doi.org/10.2139/ssrn.4949859>
- Andrews, B., Ramsey, J., & Cooper, G. F. (2018). Scoring Bayesian Networks of Mixed Variables. *International Journal of Data Science and Analytics*, 6(1), 3–18. <https://doi.org/10.1007/s41060-017-0085-7>
- Andrews, B., Ramsey, J., & Cooper, G. F. (2019). Learning High-dimensional Directed Acyclic Graphs with Mixed Data-types. *Proceedings of Machine Learning Research*, 104, 4–21.
- Arkes, H. R. (2001). Overconfidence in Judgmental Forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 495–515). Springer US.
https://doi.org/10.1007/978-0-306-47630-3_22
- Artner, S., Girven, R. S., & Bruce, J. B. (2016). Assessing the Value of Structured Analytic Techniques in the U.S. Intelligence Community: *Rand Corporation*.
https://www.rand.org/pubs/research_reports/RR1408.html
- Atanasov, P., & Himmelstein, M. (2023). Talent Spotting in Crowd Prediction. In M. Seifert (Ed.), *Judgment in Predictive Analytics* (pp. 135–184). Springer International Publishing.
https://doi.org/10.1007/978-3-031-30085-1_6

- Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Management Science*, 63(3), 691–706. <https://doi.org/10.1287/mnsc.2015.2374>
- Azzimonti, L., Corani, G., & Scutari, M. (2022). A Bayesian hierarchical score for structure learning from related data sets. *International Journal of Approximate Reasoning*, 142, 248–265. <https://doi.org/10.1016/j.ijar.2021.11.013>
- Barnes, A. (2016). Making Intelligence Analysis More Intelligent: Using Numeric Probabilities. *Intelligence and National Security*, 31(3), 327–344. <https://doi.org/10.1080/02684527.2014.994955>
- Barnett, A., Primoratz, T., Rozario, R. de, Saletta, M., Thorburn, L., & Gelder, T. van. (2021). *Analytic Rigour in Intelligence*. University of Melbourne. <https://bpb-ap-se2.wpmucdn.com/blogs.unimelb.edu.au/dist/8/401/files/2021/04/Analytic-Rigour-in-Intelligence-Approved-for-Public-Release.pdf>
- Baron, J. (1985). *Rationality and Intelligence*. Cambridge University Press.
- Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, 4(3), 265–284. <https://doi.org/10.1016/j.jarmac.2014.09.003>
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3), 800–813. <https://doi.org/10.3758/s13428-011-0081-0>
- Belliveau, J., & Yakovenko, I. (2022). Evaluating and improving the quality of survey data from panel and crowd-sourced samples: A practical guide for psychological research. *Experimental and Clinical Psychopharmacology*, 30(4), 400–408. <https://doi.org/10.1037/pha0000564>
- Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial Miscalibration*. *The Quarterly Journal of Economics*, 128(4), 1547–1584. <https://doi.org/10.1093/qje/qjt023>

- Benjamin, D. M., Hey, S. P., MacPherson, A., Hachem, Y., Smith, K. S., Zhang, S. X., Wong, S., Dolter, S., Mandel, D. R., & Kimmelman, J. (2022). Principal investigators over-optimistically forecast scientific and operational outcomes for clinical trials. *PLoS ONE*, *17*(2), e0262862. <https://doi.org/10.1371/journal.pone.0262862>
- Benson, P. G., & Önköl, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*(4), 559–573. [https://doi.org/10.1016/0169-2070\(92\)90066-I](https://doi.org/10.1016/0169-2070(92)90066-I)
- Bergh, D. van den, Clyde, M. A., Gupta, A. R. K. N., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2021). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, *53*(6), 2351–2371. <https://doi.org/10.3758/s13428-021-01552-2>
- Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5, Supplement), S2–S23. <https://doi.org/10.1016/j.amjmed.2008.01.001>
- Biais, B., Hilton, D., Mazurier, K., & Pouget, S. (2005). Judgemental Overconfidence, Self-Monitoring, and Trading Performance in an Experimental Financial Market. *The Review of Economic Studies*, *72*(2), 287–312. <https://doi.org/10.1111/j.1467-937X.2005.00333.x>
- Bier, V. (2004). Implications of the research on expert overconfidence and dependence. *Reliability Engineering & System Safety*, *85*(1), 321–329. <https://doi.org/10.1016/j.ress.2004.03.020>
- Boicu, M., Marcu, D., Tecuci, G., Kaiser, L., Uttamsingh, C., & Kalale, N. (2018). Co-Arg: Cogent Argumentation with Crowd Elicitation. *arXiv*. <http://arxiv.org/abs/1810.01541>
- Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078%253C0001:VOFEIT%253E2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078%253C0001:VOFEIT%253E2.0.CO;2)

- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 6(1), 3–5. <https://doi.org/10.1177/1745691610393980>
- Bum, C.-H., Choi, C., & Lee, K. (2018). Irrational Beliefs and Social Adaptation of Online Sports Gamblers According to Addiction Level: A Comparative Study. *Sustainability*, 10(11), 4314. <https://doi.org/10.3390/su10114314>
- Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, 11(2), 215–235. <https://doi.org/10.1007/s11409-015-9142-6>
- Canonaco, F., Gaudillo, J., Astrologo, N., Stella, F., & Acerbi, E. (2025). A guide to Bayesian networks software for structure and parameter learning, with a focus on causal discovery tools. *Frontiers in Systems Biology*, 5, 1631901. <https://doi.org/10.3389/fsysb.2025.1631901>
- Chang, W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security*, 33(3), 337–356. <https://doi.org/10.1080/02684527.2017.1400230>
- Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(509–526), 18.
- Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). *LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding*. arXiv. <https://doi.org/10.48550/arXiv.2306.14924>
- Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian Adaptive Sampling for Variable Selection and Model Averaging. *Journal of Computational and Graphical Statistics*, 20(1), 80–101. <https://doi.org/10.1198/jcgs.2010.09049>

- Collins, R. N., Mandel, D. R., Karvetski, C. W., Wu, C. M., & Nelson, J. D. (2024). The wisdom of the coherent: Improving correspondence with coherence-weighted aggregation. *Decision*, 11(1), 60–85. <https://doi.org/10.1037/dec0000211>
- Cooper, J. R. (2005). *Curing Analytic Pathologies: Pathways to Improved Intelligence Analysis*. Center for the Study of Intelligence.
- Cordova-Pozo, K., & Rouwette, E. A. J. A. (2023). Types of scenario planning and their effectiveness: A review of reviews. *Futures*, 149, 103153. <https://doi.org/10.1016/j.futures.2023.103153>
- Coulthart, S. (2016). Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency. *Intelligence and National Security*, 31(7), 933–948. <https://doi.org/10.1080/02684527.2016.1140327>
- Coulthart, S. (2017). An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques. *International Journal of Intelligence and CounterIntelligence*, 30(2), 368–391. <https://doi.org/10.1080/08850607.2016.1230706>
- Council of Europe. (2025). *The CEFR Levels—Common European Framework of Reference for Languages (CEFR)*. Common European Framework of Reference for Languages (CEFR). <https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions>
- Crawford, N. C. (2013). Civilian Death and Injury in the Iraq War, 2003-2013. *Costs of War*. <https://costsofwar.watson.brown.edu/sites/default/files/papers/Crawford-Civilian-Death-and-Injury-Iraq-War-2003-2013.pdf>
- Cross, D., Ramos, J., Mellers, B., Tetlock, P. E., & Scott, D. W. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting*, 37(3), 259–268. <https://doi.org/10.1002/for.2489>
- DeGroot, M. H., & Mortera, J. (1991). Optimal Linear Opinion Pools. *Management Science*, 37(5), 546–558.

- Demey, L., Kooi, B., & Sack, J. (2019). Logic and Probability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University.
<https://plato.stanford.edu/archives/sum2019/entries/logic-probability/>
- Dhami, M. K., Belton, I. K., & Mandel, D. R. (2019). The “analysis of competing hypotheses” in intelligence analysis. *Applied Cognitive Psychology*, 33(6), 1080–1090.
- Du, N., & Budescu, D. V. (2007). Does past volatility affect investors’ price forecasts and confidence judgements? *International Journal of Forecasting*, 23(3), 497–511.
<https://doi.org/10.1016/j.ijforecast.2007.03.003>
- Du, N., Shelton, S., & Whittington, R. (2012). Does supplementing outcome feedback with performance feedback improve probability judgments? *International Journal of Financial Research*, 3(4), 19–32. <https://doi.org/10.5430/ijfr.v3n4p19>
- Erceg, N., & Galić, Z. (2014). Overconfidence bias and conjunction fallacy in predicting outcomes of football matches. *Journal of Economic Psychology*, 42, 52–62.
<https://doi.org/10.1016/j.joep.2013.12.003>
- Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). The Maryland Scientific Methods Scale. In *Evidence-Based Crime Prevention*. Routledge.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160.
<https://doi.org/10.3758/BRM.41.4.1149>
- Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, 75, 47–52. <https://doi.org/10.1016/j.paid.2014.11.017>

- Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2), 105–109.
<https://doi.org/10.3969/j.issn.1002-0829.2014.02.009>
- Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance*, 29(3), 352–369.
[https://doi.org/10.1016/0030-5073\(82\)90250-1](https://doi.org/10.1016/0030-5073(82)90250-1)
- Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning*, 12(1), 1–19. <https://doi.org/10.1007/s11409-016-9158-6>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic Mental Models: A Brunswikian Theory of Confidence. *Psychological Review*, 98(4), 506–528.
- Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
<https://doi.org/10.1198/016214506000001437>
- González-Vallejo, C., & Bonham, A. (2007). Aligning confidence with accuracy: Revisiting the role of feedback. *Acta Psychologica*, 125(2), 221–239. <https://doi.org/10.1016/j.actpsy.2006.07.010>
- Good Judgment Inc. (2025). The first championship season. *Good Judgment*.
<https://goodjudgment.com/resources/the-superforecasters-track-record/the-first-championship-season/>
- Greenberg, S. (2020). Calibration scoring rules for practical prediction training. *arXiv*.
<http://arxiv.org/abs/1808.07501>

- Gruetzemacher, R., Lee, K. B., & Paradice, D. (2023). Calibration training for improving probabilistic judgments using an interactive app. *Futures & Foresight Science*, 6(2), e177.
<https://doi.org/10.1002/ffo2.177>
- Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, 8(3), 188–201.
<https://doi.org/10.1017/S1930297500005921>
- Hernandez, D. (2017). How our company learned to make better predictions about everything. *Harvard Business Review*. <https://hbr.org/2017/05/how-our-company-learned-to-make-better-predictions-about-everything>
- Heuer, R., Jr, R. H., & Pherson, R. (2014). *Structured Analytic Techniques for Intelligence Analysis* (2 edition). CQ Press.
- Hubbard Decision Research. (2019). *Calibrated Probability Assessments*. Hubbard Decision Research.
<https://hubbardresearch.com/shop/calibrated-probability-assessments/>
- Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning*, 4(2), 161–176.
<https://doi.org/10.1007/s11409-009-9042-8>
- Hunt Laboratory. (2019). *Hunt Challenge 2020 Brochure*. <https://cpb-ap-se2.wpmucdn.com/blogs.unimelb.edu.au/dist/8/401/files/2019/11/Hunt-Challenge-2020-Brochure.-Final.pdf>
- Hunt Laboratory, & University of Melbourne. (2020). *2020 Hunt Challenge Feedback and Review*.
- Iacus, S. M., King, G., & Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1), 1–24. <https://doi.org/10.1093/pan/mpr013>
- IARPA. (2010). *Aggregative Contingent Estimation (ACE)*. <https://www.iarpa.gov/index.php/research-programs/ace>

- IARPA. (2016). *IARPA-BAA-15-11: Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE)*. <https://www.iarpa.gov/index.php/research-programs/create/create-baa>
- Intelligence Reform and Terrorism Prevention Act of 2004 (IRTPA), Pub. L. No. 108- 458: 118 (2004).
- Ives, A. R., & Zhu, J. (2006). Statistics For Correlated Data: Phylogenies, Space, And Time. *Ecological Applications*, 16(1), 20–32. <https://doi.org/10.1890/04-0702>
- Janssen, E. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Mainhard, T., van Peppen, L. M., & van Gog, T. (2020). Psychometric properties of the Actively Open-minded Thinking scale. *Thinking Skills and Creativity*, 36, 100659. <https://doi.org/10.1016/j.tsc.2020.100659>
- Jasp Team. (2023). *JASP (Version 0.17.2.1)* [Computer software].
- Johnston, R. (2005). *Analytic culture in the US intelligence community: An ethnographic study*. Center for the Study of Intelligence, Central Intelligence Agency.
<http://catalog.hathitrust.org/api/volumes/oclc/59283799.html>
- Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, 107(2), 384–396.
<https://doi.org/10.1037/0033-295X.107.2.384>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Karvetski, C. W., & Mandel, D. (2020). Coherence of probability judgments from uncertain evidence: Does ACH help? *Judgment and Decision Making*, 15, 939–958.
- Karvetski, C. W., Mandel, D. R., & Irwin, D. (2020). Improving Probability Judgment in Intelligence Analysis: From Structured Analysis to Statistical Aggregation. *Risk Analysis*.
<https://doi.org/10.1111/risa.13443>
- Kelly, M., & Mandel, D. R. (2024). *Evaluating Commercial Confidence Calibration Training*.
<https://osf.io/yqxpw/>

- Kelly, M. O., & Mandel, D. R. (2024). The effect of calibration training on the calibration of intelligence analysts' judgments. *Applied Cognitive Psychology*, 38(5), e4236.
<https://doi.org/10.1002/acp.4236>
- Korb, K. B., Nyberg, E. P., Oshni Alvandi, A., Thakur, S., Ozmen, M., Li, Y., Pearson, R., & Nicholson, A. E. (2020). Individuals vs. BARD: Experimental Evaluation of an Online System for Structured, Collaborative Bayesian Reasoning. *Frontiers in Psychology*, 11.
<https://doi.org/10.3389/fpsyg.2020.01054>
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 107–118. <https://doi.org/10.1037/0278-7393.6.2.107>
- Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. (Second). Sage Publications.
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Edition 2). Academic Press.
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *The Journal of Applied Psychology*, 98(6), 1060–1072. <https://doi.org/10.1037/a0034156>
- Lawrence, M., Goodwin, P., O'Connor, M., & Önköl, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*, 22(3), 493–518.
<https://doi.org/10.1016/j.ijforecast.2006.03.007>
- Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26(2), 149–171. [https://doi.org/10.1016/0030-5073\(80\)90052-5](https://doi.org/10.1016/0030-5073(80)90052-5)
- Mandel, D. R. (2019). Can decision science improve intelligence analysis? In *Researching National Security Intelligence: Multidisciplinary Approaches* (pp. 117–140). Georgetown University Press.

- Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, 111(30), 10984–10989.
<https://doi.org/10.1073/pnas.1406138111>
- Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making*, 31(1), 127–137. <https://doi.org/10.1002/bdm.2055>
- Mandel, D. R., & Irwin, D. (2021). Tracking accuracy of strategic intelligence forecasts: Findings from a long-term Canadian study. *Futures & Foresight Science*, 3(3–4), e98.
<https://doi.org/10.1002/ffo2.98>
- Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making*, 13(6), Article 6.
- Mandel, D. R., & Navarrete, G. (2015). Improving Bayesian reasoning: What works and why? *Frontiers in Psychology*, 6, 1872.
- Mandel, D. R., & Tetlock, P. E. (2018). Correcting Judgment Correctives in National Security Intelligence. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.02640>
- Marcoci, A., Vercammen, A., & Burgman, M. (2019). ODNI as an analytic ombudsman: Is Intelligence Community Directive 203 up to the task? *Intelligence and National Security*, 34(2), 205–224.
<https://doi.org/10.1080/02684527.2018.1546265>
- Martin, M., & Mandel, D. R. (2025). Calibration Feedback With the Practical Scoring Rule Does Not Improve Calibration of Confidence. *Futures & Foresight Science*, 7(1), e199.
<https://doi.org/10.1002/ffo2.199>
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21(1), 1–14.
<https://doi.org/10.1037/xap0000040>

- Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine*, 173(21), 1952–1958. <https://doi.org/10.1001/jamainternmed.2013.10081>
- Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, 11(8), e12331. <https://doi.org/10.1111/spc3.12331>
- Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence Calibration in a Multiyear Geopolitical Forecasting Competition. *Management Science*, 63(11), 3552–3565. <https://doi.org/10.1287/mnsc.2016.2525>
- Muehlhauser, L. (2018). New web app for calibration training. *Open Philanthropy Blog*. <https://www.openphilanthropy.org/research/new-web-app-for-calibration-training/>
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4), 595–600. [https://doi.org/10.1175/1520-0450\(1973\)012%253C0595:ANVPOT%253E2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012%253C0595:ANVPOT%253E2.0.CO;2)
- Myung, I. J. (2000). The Importance of Complexity in Model Selection. *Journal of Mathematical Psychology*, 44(1), 190–204. <https://doi.org/10.1006/jmps.1999.1283>
- National Commission on Terrorist Attacks Upon the United States. (2004). *9/11 Commission Report Executive Summary* (Committee Report No. 206229). U.S. Department of Justice. <https://www.ojp.gov/ncjrs/virtual-library/abstracts/911-commission-report-executive-summary>
- Nicholson, A. E., Korb, K. B., Nyberg, E. P., Wybrow, M., Zukerman, I., Mascaro, S., Thakur, S., Alvandi, A. O., Riley, J., Pearson, R., Morris, S., Herrmann, M., Azad, A. K. M., Bolger, F., Hahn, U., & Lagnado, D. (2019). BARD: A structured technique for group elicitation of Bayesian networks to support analytic reasoning. *Decision Support Systems*, 34.
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2, 175–220.

- Office of the Director of National Intelligence. (2015a). *Intelligence Community Directive 203*.
<https://irp.fas.org/dni/icd/icd-203.pdf>
- Office of the Director of National Intelligence. (2015b). *Rating scale for evaluating analytic tradecraft standards with amplified guidance for evaluators (last revised on 6 November 2015)*. Office of the Director of National Intelligence.
- Önköl, D., & Muradoğlu, G. (1995). Effects of feedback on probabilistic forecasts of stock prices. *International Journal of Forecasting*, 11(2), 307–319. [https://doi.org/10.1016/0169-2070\(94\)00572-T](https://doi.org/10.1016/0169-2070(94)00572-T)
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872. <https://doi.org/10.1016/j.jesp.2009.03.009>
- Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, 76(6), 972–987. <https://doi.org/10.1037/0022-3514.76.6.972>
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–419. <https://doi.org/10.1017/S1930297500002205>
- Phelps, R. H., Halpin, S. M., Johnson, E. M., & Moses, F. L. (1980). *Implementation of Subjective Probability Estimates in Army Intelligence Procedures: A Critical Review of Research Findings* (Research Report No. 1242). U.S. Army. <https://apps.dtic.mil/sti/citations/ADA093675>
- Pherson, R. H., & Heuer, R. J. (2021). *Structured analytic techniques for intelligence analysis* (Third edition). SAGE, CQ Press.
- Pilny, A., McAninch, K., Slone, A., & Moore, K. (2024). From manual to machine: Assessing the efficacy of large language models in content analysis. *Communication Research Reports*, 41(2), 61–70. <https://doi.org/10.1080/08824096.2024.2327547>

- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 71–91. <https://doi.org/10.1111/j.1467-9868.2009.00726.x>
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>
- Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59. <https://doi.org/10.1007/BF00055564>
- 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. <https://doi.org/10.1016/j.ijforecast.2013.09.009>
- Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics*, 8(2), 1256–1280. <https://doi.org/10.1214/14-AOAS739>
- Select Committee on Intelligence. (2006). *Report of the Select Committee on Intelligence on Postwar Findings About Iraq's WMD Programs and Links to Terrorism and how They Compare with Prewar Assessments Together with Additional and Minority Views* (2005/2006; Senate Report Nos. 109–331). US Senate. <https://www.congress.gov/committee-report/109th-congress/senate-report/331/1>
- Shanks, D. R. (2006). Bayesian associative learning. *Trends in Cognitive Sciences*, 10(11), 477–478. <https://doi.org/10.1016/j.tics.2006.09.004>

- Shapira, I. (2025). Surprise, Surprise: Misperception, Hubris, and Deception in Today's Strategic Environment. *Modern War Institute*. <https://mwi.westpoint.edu/surprise-surprise-misperception-hubris-and-deception-in-todays-strategic-environment/>
- Shinitzky, H., Shemesh, Y., Leiser, D., & Gilead, M. (2024). Improving geopolitical forecasts with 100 brains and one computer. *International Journal of Forecasting*, 40(3), 958–970. <https://doi.org/10.1016/j.ijforecast.2023.08.004>
- Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making*, 18(1), 29–53. <https://doi.org/10.1002/bdm.486>
- Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3), 281–295. <https://doi.org/10.2307/3172740>
- Sinnott, R. (2019). *The Design and Development of a Cloud-based Platform Supporting Team-oriented Evidence-based Reasoning: SWARM Systems Paper*. <https://doi.org/10.24251/HICSS.2019.050>
- Stanovich, K. E., & Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence*, 11(2), 27. <https://doi.org/10.3390/jintelligence11020027>
- Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, 18(10), 1099–1104. <https://doi.org/10.1111/j.1553-2712.2011.01185.x>
- Stone, E. R., Luu, J., Costello, C. K., & Somerville, A. H. (2023). Automated calibration training for forecasters. *Journal of Behavioral Decision Making*, 36(4), e2334. <https://doi.org/10.1002/bdm.2334>
- Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, 83(2), 282–309. <https://doi.org/10.1006/obhd.2000.2910>

- Stromer-Galley, J., Rossini, P., Kenski, K., McKernan, B., Clegg, B., Folkestad, J., Østerlund, C., Schooler, L., Boichak, O., Canzonetta, J., Martey, R. M., Pavlich, C., Tsetsi, E., & McCracken, N. (2021). Flexible versus structured support for reasoning: Enhancing analytical reasoning through a flexible analytic technique. *Intelligence and National Security*, 36(2), 279–298.
<https://doi.org/10.1080/02684527.2020.1841466>
- Subbotin, V. (1996). Outcome feedback effects on under- and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes*, 66(3), 268–276.
<https://doi.org/10.1006/obhd.1996.0055>
- Syed, M., & Nelson, S. C. (2015). Guidelines for Establishing Reliability When Coding Narrative Data. *Emerging Adulthood*, 3(6), 375–387. <https://doi.org/10.1177/2167696815587648>
- Teigen, K. H. (1983). Studies in subjective probability I: Prediction of random events. *Scandinavian Journal of Psychology*, 24(1), 13–25. <https://doi.org/10.1111/j.1467-9450.1983.tb00471.x>
- Tetlock, P. E. (2006). *Expert political judgment: How good is it? How can we know?* Princeton Univ. Press.
- Tetlock, P. E., Mellers, B., Rohrbaugh, N., & Chen, E. (2014). Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in Psychological Science*, 23, 290–295. <https://doi.org/10.1177/0963721414534257>
- Thorburn, L., & Gelder, T. van. (2021). *Validity of the ODNI's IC Rating Scale as a Measure of Analytic Rigor in Intelligence Products*. <https://osf.io/cxsr8/>
- Thorburn, L., Gelder, T. van, Rozario, R. de, Kruger, A., Saletta, M., Proud, V., Young, L. J., Ganegoda, A., Barnett, A., Primoratz, T., & Thomman, S. (2020). *Quality of reasoning produced by professional and citizen intelligence analysts using the Hunt Platform*.
<https://doi.org/10.17605/OSF.IO/FDNZU>

- Thorburn, L., van Gelder, T., Shahan Brugh, C., Kim, S. M., Kolb, M., & Zulauf, B. (2022). *The IC Rating Scale as a Measure of Analytic Rigor*. Social Science Research Network.
<https://doi.org/10.2139/ssrn.4519926>
- Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012). The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions. *AAAI, Fall Symposium Series*, 6.
- United Nations. (2019). *World urbanization prospects: The 2018 revision*.
<https://doi.org/10.18356/b9e995fe-en>
- US Government. (2009). *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. Central Intelligence Agency. <https://www.cia.gov/resources/csi/static/Tradecraft-Primer-apr09.pdf>
- van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28(3), 813–826. <https://doi.org/10.3758/s13423-020-01798-5>
- van Gelder, T., De Rozario, R., & Sinnott, R. O. (2018). SWARM: Cultivating Evidence-Based Reasoning. *Computing in Science Engineering*, 20(6), 22–34. Computing in Science Engineering.
<https://doi.org/10.1109/MCSE.2018.2873860>
- van Gelder, T., Kruger, A., Thomman, S., De Rozario, R., Silver, E., Saletta, M., Barnett, A., Sinnott, R. O., Jayaputera, G. T., & Burgman, M. (2020). Improving Analytic Reasoning via Crowdsourcing and Structured Analytic Techniques. *Journal of Cognitive Engineering and Decision Making*, 14(3), 195–217. <https://doi.org/10.1177/1555343420926287>

- Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2019). A Tale of Two Sample Sources: Do Results from Online Panel Data and Conventional Data Converge? *Journal of Business and Psychology*, 34(4), 425–452. <https://doi.org/10.1007/s10869-018-9552-y>
- Wilcox, J., & Mandel, D. R. (2024). Critical review of the Analysis of Competing Hypotheses technique: Lessons for the intelligence community. *Intelligence and National Security*, 39(6), 941–962. <https://doi.org/10.1080/02684527.2024.2304934>
- Wright, G. (1982). Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica*, 52(1–2), 165–174. [https://doi.org/10.1016/0001-6918\(82\)90033-6](https://doi.org/10.1016/0001-6918(82)90033-6)
- Wright, G., & Ayton, P. (1986). Subjective Confidence in Forecasts: A Response to Fischhoff and MacGregor. *Journal of Forecasting*, 5(2), 117.
- Yates, J. F. (1990). *Judgment and decision making* (pp. xvi, 430). Prentice-Hall, Inc.
- Zacharakis, A. L., & Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, 16(4), 311–332. [https://doi.org/10.1016/S0883-9026\(99\)00052-X](https://doi.org/10.1016/S0883-9026(99)00052-X)
- Zakay, D. (1992). The influence of computerized feedback on overconfidence in knowledge. *Behaviour & Information Technology*, 11(6), 329–333. <https://doi.org/10.1080/01449299208924354>