Extracting more wisdom from the crowd

by

John Patrick McCoy

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Author .....
Department of Brain and Cognitive Sciences
26 April, 2018

Signature redacted

Certified by .....
Dražen Prelec
Digital Equipment Corporation Leaders for
Global Operations Professor of Management
Thesis Supervisor

Signature redacted

Certified by .....
Joshua B. Tenenbaum
Professor of Computational Cognitive Science
Thesis Supervisor

Signature redacted

Accepted by .....
Matthew Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

Extracting more wisdom from the crowd

by

John Patrick McCoy

Submitted to the Department of Brain and Cognitive Sciences
on 26 April, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

In many situations, from economists predicting unemployment rates to chemists estimating fuel safety, individuals have differing opinions or predictions. We consider the wisdom-of-the-crowd problem of aggregating the judgments of multiple individuals on a single question, when no outside information about their competence is available. Many standard methods select the most popular answer, after correcting for variations in confidence. Using a formal model, we prove that any such method can fail even if based on perfect Bayesian estimates of individual confidence, or, more generally, on Bayesian posterior probabilities. Our model suggests a new method for aggregating opinions: select the answer that is more popular than people predict. We derive theoretical conditions under which this new method is guaranteed to work, and generalize it to questions with more than two possible answers. We conduct empirical tests in which respondents are asked for both their own answer to some question and their prediction about the distribution of answers given by other people, and show that our new method outperforms majority and confidence-weighted voting in a range of domains including geography and trivia questions, laypeople and professionals judging art prices, and dermatologists evaluating skin lesions. We develop and evaluate a probabilistic generative model for crowd wisdom, including applying it across questions to determine individual respondent expertise and comparing it to various Bayesian hierarchical models. We extend our new crowd wisdom method to operate on domains where the answer space is unknown in advance, by having respondents predict the most common answers given by others, and discuss performance on a cognitive reflection test as a case study of this extension.

Thesis Supervisor: Dražen Prelec
Title: Digital Equipment Corporation Leaders for
Global Operations Professor of Management

Thesis Supervisor: Joshua B. Tenenbaum
Title: Professor of Computational Cognitive Science

Acknowledgments

It's a tremendous pleasure to thank many people.

First, my advisors Dražen Prelec and Josh Tenenbaum, for innumerable things. Through their example, counsel, and support they've contributed to not only my intellectual development, but also my personal development. Dražen gave me the freedom to read and explore, but was also there whenever I needed him. If I've acquired a small fraction of his wisdom and his taste this will have been well worth it. Josh's vision and inspiration pervade how I think about cognitive science, and being a scientist and colleague. He's cared about me and my career from the beginning of my MIT journey, for which I'm extremely grateful.

Laura Schulz for serving as my committee chair, but also for her encouragement for a line of research that doesn't appear in this thesis, but that I expect will influence much of my future thinking.

Shane Frederick for serving on my committee, useful feedback, and making it so much fun to talk and think about judgment and decision making. I hope we'll do lots of good work together.

Sebastian Seung for his collaboration developing the surprisingly popular answer, his smarts and his standards.

Danica Mijovic-Prelec, Alex Huang, and Murad Alam for help designing and conducting the skin lesions study, and Danielle Suh for help designing and conducting the art studies.

The participants in my research, for their time and effort.

The host of people who took the time to give me useful feedback about how to best present this work towards the end of the process, including Artem Timoshenko, Shuyi Yu, Sachin Banker, Colleen Giblin, Carey Morewedge, Tony Ke, Dean Eckles, Juanjuan Zhang, Sharmila Chatterjee, Catherine Tucker, John Hauser, Birger Wernerfelt, Pedro Tsividis, Tobi Gerstenberg, Ilker Yildrim, Kevin Smith, Tomer Ullman, Shane Frederick, Josh Tenenbaum, and Dražen Prelec.

My CoCoSci compatriots, broadly construed, for their friendship, feedback, knowl-

edge, and inspiration: Tobias Gerstenberg, Ilker Yildirim, Kevin Smith, Sydney Levine, Jon Malmaud, Max Kleiman-Weiner, Max Siegel, Pedro Tsividis, Kelsey Allen, Josh Rule, Kevin Ellis, Mario Belledonne, Amir Soltani, Eliza Kosoy, Vikash Mansinghka, Cameron Freer, Chris Baker, Eyal Dechter, Peter Krafft, Tim O'Donnell, Yibiao Zhao, Julian Jara Ettinger, David Reshef, Roger Grosse, Peter Battaglia, Jessica Hamrick, Leon Bergen, Josh Hartshorne, Sam Gershman, Rus Salakhutdinov, Dan Roy, Steve Piantadosi, Virginia Savova, Noah Goodman, Mike Frank, Ed Vul, Frank Jäkel, Liz Bonawitz, Hyo Gweon, Sam Zimmerman, and Yarden Katz.

Ditto for the folk associated with the Sloan Neuroeconomics Lab: Danica Mijovic-Prelec, Catherine Holland, Ryan Hauser, Ursa Bernardic, Dae Houlihan, Sonja Radas, Alex Huang, Chris Long, Kaustubh Patil, Josh Manning, and Derek Dunfield.

The SPADE team for the chance to work together to improve crowd wisdom at scale, Sonja Radas for the chance to work together to improve crowd wisdom for market predictions, and Laurie Paul for the chance to work together on modal prospection (but really for the *de re* and *de se*).

The graduate students, postdocs, and faculty of BCS and Sloan Marketing. Many students are lucky to be part of a wonderful department community, I was lucky enough to get two.

The B-Lab folk for useful feedback.

Denise Heintze, Tobi Momoh, Julianne Ormerod, and Jason Clinkscales for cheerful support and removing obstacles.

Fulbright, IARPA, Citibank, and Google for keeping me in lucre.

Michael Henning and David Spurrett for getting me started, and sending me on my way.

Hugh Pastoll, Michael Meadon, Wayne Christensen, Andries Gouws, John Collier, Deepak Mistry, Julia Clare, Jochen Zeller, Heike Tappe, Yolanda Hordyk, Andrew Dellis, Marc Robson and Eva Jackson for friendship and helping me realise in those Durban cappuccino days that cognitive science is what I wanted to do.

David Wingate for his early mentoring and collaboration at CoCoSci.

Tomer Ullman for some kind of mind-meld.

Andreas Stuhlmüller for thoughtful questions, thoughtful answers, and trying to do good.

Brenden Lake for his friendship while demonstrating how to make steady progress on big questions.

Nell Putnam-Farr for excellent coffee breaks and helpful insight.

Carey Morewedge for sage advice and taking the time.

Ben Murrell for friendship, high intellectual standards, and our early attempts at cognitive science.

1st Pmb. and P.L.T.U. for helping build the neccessary character.

The magical Plymouth Street - what are the odds that three houses on one street would temporarily enclose some of my now closest friends?[1]

I count myself in nothing else so happy / As in a soul remembering my good friends.[2]

Marjorie Friedman, for being there.

My family, for everything.

[1] We leave this as an exercise for the reader. Hint: take the correlation structure into account, and choose an appropriate prior.

[2] Richard II 2.3.47-48.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Crowd wisdom and information aggregation

1.1 The promise of crowd wisdom

Consider the following four scenarios:

(a) Chemists are asked to estimate the relative safety of new proposed fuels for automobiles by considering both their physical properties and potential accidents.

(b) Economists and political scientists are asked to produce forecasts about economic and political events in various countries, for example the probable outcome of an ongoing regional conflict.

(c) Doctors are asked to diagnose whether a possibly cancerous tumor is benign or malignant.

(d) Consumers are asked about how likely they are to make use of various products in the future which have not yet been introduced into the market.

It is a truism that groups of people, particularly groups of experts, are a better source of answers to questions like the above than a single individual[1] and indeed groups

[1]The quote from the cover of the bestseller "The Wisdom of Crowds" (Surowiecki, 2005) is typical of this enthusiasm: "Large groups of people are smarter than an elite few, no matter how brilliant

of people make these kinds of judgments and decisions every day. The impressive successes of such (sometimes large) groups of people are catalogued in "The Wisdom of Crowds" (Surowiecki, 2005) and "Infotopia" (Sunstein, 2006), both of which forcefully argue for the role of the dispersed judgments of the many in policy making. Recent applications of crowd wisdom include political and economic forecasting (Budescu and Chen, 2014; Mellers et al., 2014), corporate decision making (Bonabeau, 2009), and healthcare (Brabham et al., 2014). Crowds have similarly been used to evaluate nuclear safety (Cooke and Goossens, 2008), public policy(Morgan, 2014) the quality of chemical probes, (Oprea et al., 2009) and even possible responses to a restless volcano (Aspinall, 2010).

1.1.1 The challenge posed by this promise

The search for crowd wisdom, however raises a challenge: how can this wisdom best be extracted from individual judgments? That is, how should the judgments of individuals in a group be aggregated, and how can expertise be identified?

The focus of this thesis is on how to aggregate answers from a number of individuals who have independently given their answer to some question or their opinion about some issue. This enables one to aggregate the opinions of large, dispersed groups of individuals who give their opinions at different times and to additionally avoid some of the potentially negative effects of discussion (Lorenz et al., 2011), including group polarisation (Myers and Lamm, 1976; Sunstein, 2002; Isenberg, 1986). [2]More specifically, this thesis answers the the question of how to aggregate information from a number of individuals for a single, unique question where no outside information is available. Some statistical methods, such as the existing Bayesian hierarchical models we will discuss, require data from multiple questions, either because this is required to learn respondent-level parameters or because an individual's historical accuracy is

- better at solving problems, fostering innovation, coming to wise decisions, even predicting the future."

[2]This focus does not imply, of course, that group discussion should always be avoided or ignored; see (Mellers et al., 2014) for a recent success that included having individuals in the group interact.

used as model input thus limiting the application of such methods.[3] Thus, while we will extend the crowd wisdom models that we develop to apply across data from multiple questions (Chapter 3) which has various advantages when this data is available, we will primarily focus on models and methods that do not require this.

1.1.2 The scope of this thesis

Beyond the specific questions described above, we also restrict our scope in a number of other ways. First, we only consider aggregating group judgments or beliefs, not group preferences. Hanson (2013) gives one possible defense of this restriction to beliefs. Second, we assume that respondents are answering honestly, rather than attempting (either individually or through collusion) to engage in deception or force a particular answer to be chosen. As we will discuss, however, there are methods to incentivize respondents to answer honestly that use the same inputs as our aggregation methods, and many of the empirical studies we discuss are incentive compatible. Third, the thesis discusses how to aggregate answers to categorical questions with a finite number of answers, rather than questions where the answer is a continuous quantity. We leave extending these methods to domains with continuous answers to future work, but remark that one can go a long way towards aggregating continous quantities by asking a series of binary questions about whether the continuous quantity is over or under some point, or by discretizing the continuous quantitiy (not necessarily uniformly) and then treating it as a categorical variable.

1.2 Previous approaches to aggregating information from the crowd

There is now a large literature suggesting answers to these question of how to extract crowd wisdom. There are both mathematical methods for aggregation (for reviews see Cooke (1991); Clemen and Winkler (1999, 2007); Ouchi (2004)) and behavioral

[3]Even in cases where respondents have answered other questions, it is often difficult to assess respondent's historical performance (Tetlock, 2005; Denrell and Fang, 2010).

methods (for reviews see Cooke (1991); Clemen and Winkler (1999)), which in turn draw on work on decision making in groups and organizations (for reviews see Kerr and Tindale (2004); Tindale and Kluwe (2015); Larrick et al. (2011)).

1.2.1 Voting and averaging

In broad strokes, the prevalent answer to the question of how to aggregate crowd wisdom has changed little since Galton's demonstration that the average estimate of a group of county fairgoers juding the weight of an ox was remarkably accurate (Galton, 1907c,a,b): 'epistemic democracy', or, more specifically, select the average opinion. Standard methods of information aggregation, and the focus of much research, involve essentially selecting the modal opinion: majority vote in the case of categorical judgments (Austen-Smith and Banks, 1996; Keuschnigg and Ganser, 2016; DeGroot, 1974; Grofman et al., 1983; Hastie and Kameda, 2005; Ladha, 1992) or selecting the mean opinion in the case of continous judgments, including judgments of probabilities (Ashton and Ashton, 1985; Makridakis and Winkler, 1983). Alternatives to selecting the arithmetic mean include computing trimmed means (Jose et al., 2013), averaging quantiles rather than probabilities (Lichtendahl Jr et al., 2013), selecting the median of the cumulative distribution function (Hora et al., 2013), and selecting the geometric average (Genest et al., 1984).

1.2.1.1 Diversion: axiomatic justification for averaging methods

How are the averaging methods discussed above theoretically justified? Axiomatic approaches to aggregation consider how to combine probability distributions so that the resulting distribution obeys certain attractive properties, and so derive restrictions on the functional form of the resulting distribution (see Genest and Zidek (1986) and French and Insua (2000) for reviews). The axiomatic approach has mostly considered how to combine probability means given a weight associated with each mean, for example a probability judgment from an individual and an expertise weight of some kind associated with the individual. The two best known such solutions are averag-

ing methods: the linear opinion pool, which uses the weighted sum of the probability means, i.e. confidence-weighted voting, and the logarithmic opinion pool, which uses the (appropriately normalized) product of the means with each mean raised to its associated weight. The linear opinion pool posesses a desired property called the marginalization property and the logarithmic opinion pool fulfills the principle of external Bayesianity. Whilst much attention has been paid to how to use the weights in question, considerably less attention has been paid to principled ways to obtain the weights,. Beyond methods that apply to single questions, the most widely used method for aggregating expert opinions in this tradition, other than simply giving equal weight to all experts, is Cooke's classical model (Cooke, 1991), which uses a linear opinion pool with weights that are obtained from each individual's performance answering questions about seed variables for which the true outcome is known. Each individual's performance when estimating the seed variables is used to calculate calibration and informativeness scores that are together used to calculate the weight assigned to each expert. Cooke and Goossens (2008) review the performance of the classical model to aggregate opinions made by 45 expert panels over many domains.

1.2.1.2 Disadvantages of voting and averaging

Unfortunately, as this subsection discusses, simply selecting the majority or average opinion is not always successful for a variety of reasons, although it has the advantage that it is simple to apply and works well in some circumstances. For an excellent extended survey of potential drawbacks of these methods see (Sunstein, 2006). The major weakness of such democratic averaging is that it does not take into account differences in individual competence, expertise or access to information, and eliciting confidences from respondents does not adequately capture these differences (Koriat, 2012). Each individual has the same impact on the collective outcome, or, in confidence weighted averaging, is free to determine their own level of impact. Such methods are successful at identifying the correct answer only in so-called 'kind' environments (Hertwig, 2012), where individuals with the best information are either the most confident or the most numerous. They fail in 'wicked' environments, for

example, when many people in the group have misleading intuitions (Simmons et al., 2011) or when unknowedgeable or unskilled people give extreme forecasts (Evgeniou et al., 2013).

A standard justification for trusting the average or majority opinion of large groups of people is the Condorcet Jury Theorem (Condorcet, 1785), which, loosely, states that if a group of people answer a binary question and each individual answers independently and has more than 50% chance of being correct then the probability that majority vote is correct tends towards 100% as the group size increases towards infinity. There is now a large literature concerning extensions and variations of Concordet's original idea; see (Sunstein, 2006) for a survey of this literature. Whilst this kind of argument is normally used to motivate selecting the average or consensus answer, the darker side of this theorem follows immediately: if the majority of the group have less than a 50% chance of giving the correct answer, then the probability that majority vote is correct falls towards zero as the group size increases. In situations, therefore, in which many people have limited or misleading information or intuitions, selecting the average answer may lead to an incorrect conclusion.

If each individual's answer accords with the information that they have available, the majority answer will be biased towards informaton which is most widely available, rather than information which is most diagnostic of the truth. In particular, democracy may overweight shallow information that is widely available and neglect important information that is not widely dispersed, for example because it is new or hard to understand. Furthermore, majority voting (or averaging) considers only how people vote in the actual state of the world, without taking into account how they might vote in other counterfactual world states. This matters, for example, in situations in which most people are biased to vote in a particular direction in all possible world states. For example, people may (for sensible reasons) be predisposed to say that a tumor is malignant such that when the tumor is benign 65% of the responding doctors say 'malignant' but if we could somehow probe votes for the same tumor in the counterfactual world where it actually is malignant 90% of the doctors would have said it is malignant. Lastly, selecting the majority or average opinion

does not account for the prior over possible states of the world, independent of the answers of respondents. That is, there may be some state of the world which is ex ante extremely unlikely but which is such that it results in information dispersed in such a way throughout the group that slightly favors it over some other state which is much more likely ex ante. For all of these (related) reasons, the majority or average opinion can result in the incorrect answer.

We should, of course, distinguish between cases where selecting the average opinion arrives at the incorrect answer because everyone in the group was simply missing some critical piece of information or the intrinsic irreducible uncertainty of the question is simply too large (Tetlock, 2005), from situations where there were individuals in the group who had sufficient information to answer correctly but selecting the average answer did not weigh their opinions highly enough. It is the latter case where this thesis will hope to provide a better alternative.

1.2.2 Market-based aggregation mechanisms

Whilst the mathematical methods for aggregation just surveyed have a long history, more recently the internet has encouraged aggregation mechanisms that allow many dispersed people interacting repeatedly to make predictions or form judgments. The most widely used are prediction markets, in which people trade securities that pay off depending on the outcome of a specified event, with the market price of the security interpreted as the best collective estimate of the current probability of the event (Anders and Batchelder, 2012; Wolfers and Zitzewitz, 2004; Arrow et al., 2008). One major application historically of prediction markets is election forecasting (Berg et al., 2008), although they have also been tested, for example, in predicting NFL games (for which they performed similarly to an expert opinion pool) (Chen et al., 2005), and in numerous other contexts (Wolfers and Zitzewitz, 2004). Prediction markets are used within many organisations, especially businesses (Thompson, 2012; Waitz and Mild, 2013), including large-scale prediction markets of various kinds at Google and Ford Motor Company (Cowgill and Zitzewitz, 2015), and markets for sales forecasts at Hewlett Packard (Plott and Chen, 2002). Companies have also run prediction

markets with external traders, for example to forecast Google's market capitalization prior to their initial public offering (Berg et al., 2009). Beyond prediction markets, other market based mechanisms include preference markets (Dahan et al., 2010) and securities trading of concepts (Dahan et al., 2011) which are intended to aid product development by determining how consumers respond to different product concepts. Other approaches in this literature have combined a market and a coordination game to eliminate public knowledge biases (Chen et al., 2004), and, more recently, exploited parimutuel betting to forecast sales at Intel (Plott et al., 2014) and to predict box office revenue Court et al. (2018).

1.2.2.1 Advantages and disadvantages of prediction markets

Prediction markets provide incentives for people to make accurate judgments by paying people depending on the outcome of their stock trades, and furthermore allow people who purchase more stocks (reflecting how confident they are) to have a larger impact on the final result. Importantly, prediction markets elicit from people not only what information they have, but also implicitly ask people to reflect on how this information or insight is distributed amongst the population and to what extent it is already accounted for in the current stock price. However, prediction markets also have several important disadvantages (Croxson, 2011). Primarily, they are limited to predictions about events where it is possible to describe the the outcomes with a precise contract. They thus cannot be applied to forecasts about counterfactuals (for example, to forecast the results of several possible mutually exclusive public policies) or about events that resolve in the far future (for example, to forecast whether taking some action now will have a particular consequence in twenty years). Beyond this limitation, markets also require sufficient liquidity to enable accurate predictions (Sunder, 1992; Ho and Chen, 2007), and their output is by definition available to all market participants, which may be undesirable in some settings (Croxson, 2011). They can suffer from the result of individuals interacting such as bubbles, information cascades and contagion from wishful thinking (Seybert and Bloomfield, 2009; Anderson and Holt, 1997; Scharfstein and Stein, 1990).

1.2.3 Bayesian aggregation mechanisms

While the focus of this thesis is on aggregating answers to single, unique questions for which no outside information is available, Bayesian models of aggregation can be useful when the aggregator has a strong prior belief about the answer, or, more commonly, has data from respondents answering multiple questions. In Chapter 3, we develop a Bayesian probabilistic generative model that can be applied to single questions but also make use of data from respondents answering multiple questions and so we briefly discuss here previous Bayesian models of aggregation.

The idea behind the Bayesian approach to aggregating expert opinions is that the aggregator has a prior belief over the variable of interest and updates this prior with respect to a likelihood function associated with the answers of experts about the variable of interest Winkler (1968); Morris (1977), for example the aggregator may assume a uniform prior and treat each expert's opinion as a draw from a Beta distribution. Clemen and Winkler (1990) compare Bayesian models for determining the value of an indicator variable where each expert gives the probability that the indicator variable is on. Bayesian models of aggregation also exist where the variable of interest is a continuous probability, with most assuming a normal distributions around the true probability (Winkler, 1981; Lipscomb et al., 1998).

More recently, a number of Bayesian hierarchical models have been developed in computer science for aggregating information, including both answers to categorical and continuous questions (Oravecz et al., 2013; Lee and Danileiko, 2014; Lee et al., 2012, 2011b; Yi et al., 2010a). Such Bayesian hierarchical models can only be applied to multiple questions and operate by learning the value of latent parameters, for example respondent expertise, to better determine the aggregate answer. These kinds of models will be discussed more extensively in Chapter 3. Bayesian hierarchical models have also been developed for aggregating richer kinds of information, including aggregating rank order information from human memory (Steyvers et al., 2009), an extension that incorporates a paramter for individual expertise parameter (Lee et al., 2011a), and a model for aggregating answers to the traveling salesperson problem (Yi

et al., 2012).

1.2.4 Summary of previous approaches to aggregation

Leaving aside methods that rely on respondents answering multiple questions, we have considered two major classes of aggregation mechanisms. First, statistical methods that essentially extract the average or consensus answer in the group by selecting the modal opinion or by computing some kind of average. Whilst such methods have the advantage that they are democratic, easy to apply and preserve independence of judgment, they can fail in environments where only a minority of experts have the necessary information or insight, and where respondents do not realize this. An alternative approach is to aggregate opinions using a market-based mechanism such as prediction markets. While such methods potentially allow expert opinion to be highly weighted, they are limited in that they only apply to judgments that are verifiable.

1.3 Thesis structure

In Chapter 2 (joint work with Sebastian Seung and Dražen Prelec), we develop a new method of extracting crowd wisdom, which we call selecting the 'surprisingly popular' answer. We propose a formal model of how Bayesian respondents vote and predict the votes of other people, and show, under this model and the assumption of an infinite sample of Bayesian respondents, that our new method is superior to standard methods. We additionally show that across a range of empirical domains our new method delivers good performance in practice.

In Chapter 3 (joint work with Dražen Prelec), we treat the crowd wisdom problem as one of statistical inference. We develop a probabilistic generative model that builds on the model we present in Chapter 2 to overcome various limitations of selecting the surprisingly popular answer. Advantages of the probabilsitic generative model that we develop include producing a complete posterior distribution over possible answers (rather than simply which answer is more likely) and an ability to infer individual respondent expertise by taking into account information across multiple questions

that respondents have answered.

In Chapter 4 (joint work with Shane Frederick and Dražen Prelec), we extend our aggregation methods to the case where the space of answers is unknown in advance, and examine the answers and predictions of others that people give on a cognitive reflection test as a case study.

Finally, we conclude the thesis with some brief remarks.

Chapter 2

A solution to the single question crowd wisdom problem

This chapter consists largely of material from Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532-535 and its supplementary material, although includes some additional remarks and figures.

2.1 Introduction

Once considered provocative (Galton, 1907c,a,b), the notion that the wisdom-of-the-crowd is superior to any individual has itself become a piece of crowd wisdom, fueling speculation that online voting may soon put credentialed experts out of business (Sunstein, 2006; Surowiecki, 2005). Algorithms for extracting wisdom from the crowd are typically based on a democratic voting procedure. They are simple to apply and preserve independence of personal judgment (Lorenz et al., 2011). However, democratic methods have serious limitations. They are biased for shallow, lowest-common-denominator information, at the expense of novel or specialized knowledge that is not widely shared (Chen et al., 2004; Simmons et al., 2011). Adjustments based on measuring confidence do not solve this problem reliably (Hertwig, 2012). Here we propose the following alternative to a democratic vote: Select the answer that is *more*

popular than people predict. We prove that this principle yields the best answer under reasonable assumptions about voter behavior, while the standard 'most popular' or 'most confident' principles fail under exactly those same assumptions. Like traditional voting, the principle accepts unique problems, such as panel decisions about scientific or artistic merit, and legal or historical disputes. The application domain is thus distinct from that covered by machine learning and psychometric methods, which require data acrosss multiple questions (Batchelder and Romney, 1988; Lee et al., 2012; Yi et al., 2012; Lee and Danileiko, 2014; Anders and Batchelder, 2012; Oravecz et al., 2013; Freund and Schapire, 1997) .

2.2 The surprisingly popular answer

To motivate our solution, imagine that you have no knowledge of U.S. geography and are confronted with questions like:

(P) Philadelphia is the capital of Pennsylvania: Yes or No ?

(C) Columbia is the capital of South Carolina: Yes or No ?

You pose them to many people, hoping that majority opinion will be correct. This works for (C), but most people endorse the incorrect answer Yes for (P), as shown by the data in Figure 2-1(a, b). Most respondents may only recall that Philadelphia is a large, historically significant city in Pennsylvania, and conclude that it is the capital (Goldstein and Gigerenzer, 2002). The minority who vote No likely possess an additional piece of evidence, that the capital is Harrisburg. A large panel will surely include such individuals. The failure of majority opinion cannot be blamed on an uninformed panel or flawed reasoning, but represents a defect in the voting method itself.

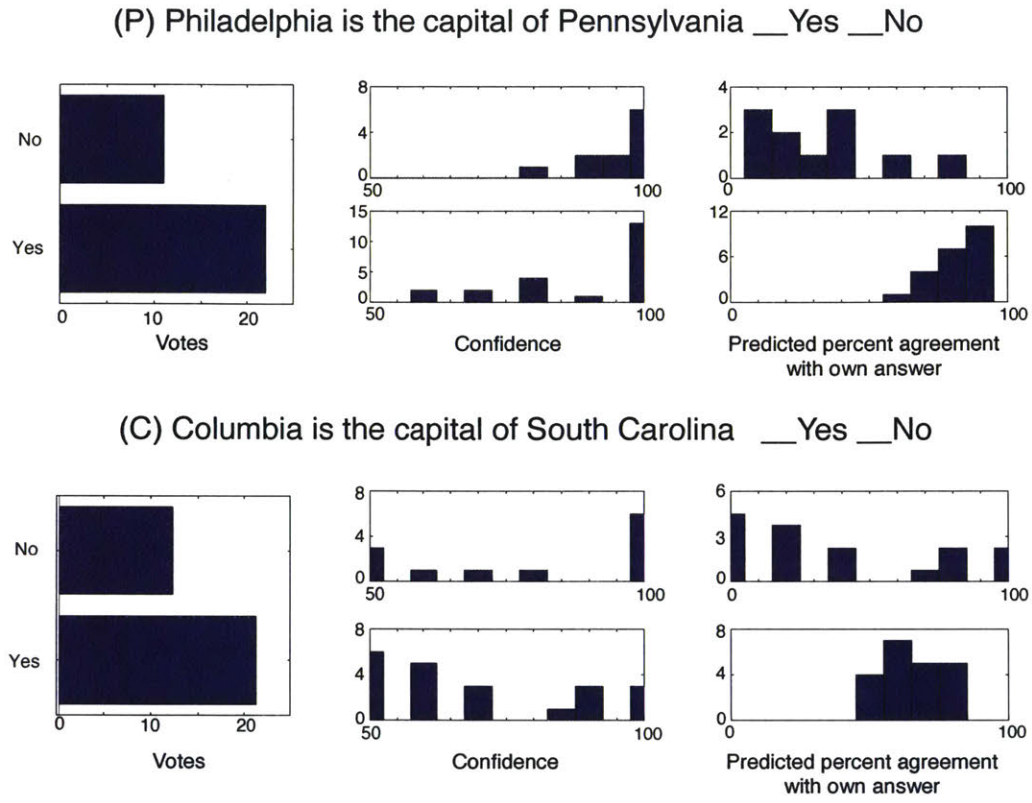


Figure 2-1: Two example questions from Study 1c, described in text. **(a)** Majority opinion is incorrect for question (P). **(b)** Majority opinion is correct for question (C). **(c)** and **(d)**. Respondents give their confidence that their answer is correct from 50% (chance) to 100% (certainty). Weighting votes by confidence does not change majority opinion, since respondents voting for both answers are roughly equally confident. **(e)** Respondents predict the frequency of *Yes* votes, shown as estimated percent agreement with their own answer. Those answering *Yes* believe that most others will agree with them, while those answering *No* believe that most others will disagree. The surprisingly popular answer discounts the more predictable votes, reversing the incorrect majority verdict in (P). **(f)** The predictions are roughly symmetric, and so the surprisingly popular answer does not overturn the correct majority verdict in (C).

A standard response to this problem is to weight votes by confidence. For binary questions, confidence c implies a subjective probability c that a respondent's vote is

correct and $1 - c$ that it is incorrect. Probabilities may be averaged linearly or non-linearly, producing confidence-weighted voting algorithms (Cooke, 1991). However, these succeed only if correct votes are accompanied by sufficiently greater confidence, which is neither the case for (P) and (C), nor more generally (Koriat, 2012). As shown by Figure 2-1(c, d), confidences associated with Yes and No votes are roughly similar and do not override the incorrect majority in (P).

Here we propose an alternative algorithm that asks respondents to predict the distribution of other people's answers to the question. The intuition underlying the algorithm is as follows. Imagine that there are two possible worlds, the actual one in which Philadelphia is not the capital of Pennsylvania, and the counterfactual one in which Philadelphia is the capital. It is plausible that in the actual world fewer people will vote Yes than in the counterfactual world. After all, in the actual world where Philadelphia is not the capital, some fraction of respondents presumably know the actual capital city and so would vote No. This can be formalized by the toss of a biased coin where, say, the coin comes up Yes 60% of the time in the actual world and 90% of the time in the counterfactual world. Majority opinion favors Yes in both worlds. People know these coin biases (the vote frequencies in the different world states) but they do not know which world is actual. Consequently, their predicted frequency of Yes votes will be between 60% and 90%, say 80% on average. However, the actual frequency of Yes votes will converge to 60% since Philadelphia is not the capital and No will be the surprisingly popular, and correct, answer. Similarly, we can consider the counterfactual world where Philadelphia is the capital city, and ask what would happen if this world was actual. In this case, the frequency of Yes votes would converge to 90% and Yes would be the surprisingly popular, answer and correct in this case. Hence, the surprisingly popular answer selects the correct answer both in the actual world, and in the counterfactual world.

We refer to this selection principle as the 'surprisingly popular' (SP) algorithm, and define it rigorously in the next section (Theorem 2). In problem (P), the data shows that respondents voting Yes believe that almost everyone will agree with them, and respondents voting No also tend to believe that most people will vote *Yes* (Figure

2-1e). On average, the predicted percentage of Yes votes is high, causing the actual percentage for Yes to underperform relative to predictions. Therefore the surprisingly popular answer is No, which is correct. In (C), by contrast, predictions of Yes votes fall short of actual Yes votes. The surprisingly popular answer agrees with the popular answer, and the majority verdict is correct (Figure 2-1f).

Could an equally valid algorithm be constructed using respondents' confidences? Assume that respondents know the prior world probabilities and coin biases. Each respondent observes the result of her private coin toss, and computes her confidence by applying Bayes' rule. The hypothesized algorithm would need to identify the actual coin from a large sample of reported confidences. Figure 2-2 proves by counterexample that no such algorithm exists. It shows how identical distributions of confidences can arise for two different biased coin problems, one where the correct answer is *Yes* and one where the correct answer is *No*. Theorem 1 generalizes this counterexample. Admittedly, real people may not conform to the idealized Bayesian model. Our point is that if methods based on posterior probabilities (votes and confidences) fail for ideal respondents, they are likely to fail for real respondents.

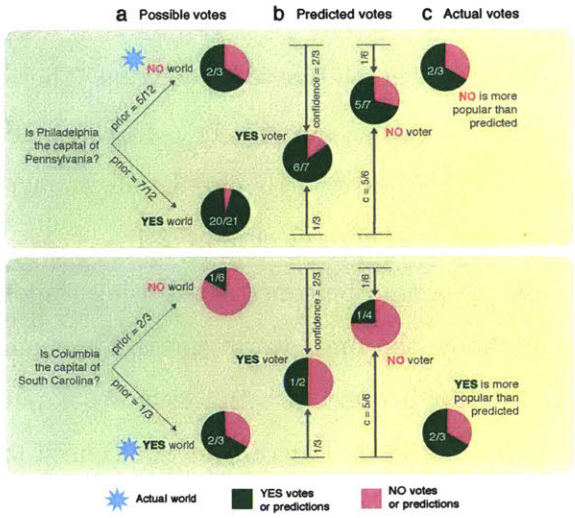


Figure 2-2: Why "surprisingly popular" answers should be correct and confidence-weighted voting is insufficient, illustrated by simple models of Philadelphia and Columbia questions with Bayesian respondents. **(a)** The correct answer is more popular in the actual world than in the counterfactual world. **(b)** Respondents' vote predictions interpolate between the two possible worlds. In both models, interpolation is illustrated by a voter with 2/3 confidence in *Yes* and a voter with 5/6 confidence in *No*. The prediction of the *Yes* voter is closer to the percentage in the *Yes* world, and the prediction of the *No* voter is closer to the percentage in the *No* world. Both predictions lie between actual and counterfactual percentages. **(c)** follows from **(a)** and **(b)**. The correct answer is the one that is more popular in the actual world than predicted — the "surprisingly popular" answer. The example also proves that any algorithm based on votes and confidences can fail even with ideal Bayesian respondents. The two questions have different correct answers, while the actual vote splits and confidences are the same. Numerical confidences were constructed from a Bayesian model in which the actual world is drawn according to a prior probability distribution, representing evidence that is common knowledge among all respondents. A respondent's vote is generated by tossing the coin corresponding to the actual world. A respondent uses their vote as private evidence to update the prior into posterior probabilities via Bayes' rule. For example, a *Yes* voter for Philadelphia would compute posterior probability of $2/3 = (7/12) \times (20/21)/((7/12) \times (20/21) + (5/12) \times (2/3))$ that *Yes* is correct.

By comparison, the SP algorithm has a theoretical guarantee, that it always selects the best answer in light of available evidence (Theorem 2). Theorem 3 extends the approach to multiple choice questions, and shows how vote predictions can identify respondents that place highest probability on the correct answer. These results are based on a common theoretical model that generalizes the biased coin example to multiple, many-sided coins.

2.3 Possible world model and theoretical results

The formal model builds on the biased coin example, generalizing to m coins, each coin having n possible sides. In part, this is a standard normative account of how individuals should make inferences about hypotheses (coins) from data (toss outomes). The additional assumption is that individuals take these inferences one step further, and compute correct expectations of tosses observed by others. We will assume an infinite sample of respondents.

A more complete Bayesian model would have parameters for respondent errors and biases, and would also deal with the finite sample issue. We will consider such a model in Chapter 3. However, it is important to first understand what can be deduced from different types of input in the ideal case. This sets boundaries on what one might expect to achieve with richer models.

We begin with a negative result, Theorem 1, that an infinite sample of correctly computed posterior probabilities over coins is compatible with any possible coin (i.e. answer) being correct. This reveals a limitation of methods based on posterior probabilities, such as votes and confidences.

We then show how to determine the correct answer in three increasingly complex settings: (1) $m = 2, n \geq 2$, (2) $m = n > 2$, (3) $m, n \geq 2$. In particular, Theorem 2 proves that the surprisingly popular algorithm for binary questions described in the main text is valid.

Extensions to multiple choice questions ($m > 2$ coins) rely on a key Lemma, which Theorem 3 applies to the $m = n > 2$ case. We then consider the fully general m, n

problem, and indicate, without formal proof, how the correct answer can be derived if in addition to vote predictions one also elicits posterior probabilities (whose elicitation is not required for Theorems 2 and 3).

The results presented here which justifying choosing the surprisingly popular answer assume ideal Bayesian respondents. However, the biased coin argument presented in the main text remains valid even with certain departures from Bayesian rationality. For example, respondents might simplify the prediction task by predicting the vote split in the world that they think is more likely, and ignoring the possibility of the less likely world. Then, those voting for the correct answer will make accurate predictions, while those voting for the wrong answer will underestimate the vote for the correct answer. The average predicted vote for the correct answer will again underestimate the actual vote, confirming the surprisingly popular principle.

2.3.1 Possible world model

The model extends the biased coin example in two ways. First, we generalize to an arbitrary number m of possible worlds (each containing a possible coin). One of the worlds is actual, the rest are counterfactual. We identify worlds with possible answers to a multiple choice question. Uncertainty about the actual world, i.e., the correct answer, is modeled by a random variable taking on values in the set $\{a_1, \ldots, a_m\}$ of m possible answers to a multiple-choice question. Second, we distinguish between a respondent's vote for a particular answer and the evidence on which that vote is based. The evidence respondent r possesses is summarized by a private 'signal' S^r, which is a random variable taking on categorical values in the set $\{s_1, \ldots, s_n\}$. A respondent's vote V^r is given by a function $V^r = V(S^r)$ that maps signals to votes $V^r \in \{v_1, \ldots, v_m\}$ for the m possible answers.

Conditional on world a_i, signals of different respondents are independent, identically distributed with probabilities $p(s_k|a_i)$. Therefore, all differences in knowledge are captured by signals. The prior $p(a_i)$ gives probabilities consistent with the evidence that is common knowledge among all respondents. For problem (P) discussed in the main text, common knowledge might be that Philadelphia is a large city. Ideal

respondents know the joint distribution $p(s_k, a_i)$, which defines the possible world model (to avoid degeneracies, we assume $p(a_i) > 0, p(s_k) > 0$). However, they do not know which a_i is the correct answer a_{i^*}, and nor do they know the actual distribution of received signals. In terms of the coin example, they know which coins are possible and the properties of each coin, but they do not know which coin is actually being used.

Respondents have two types of beliefs, both computed from their received signal s_k and the joint distribution $p(s_k, a_i)$. Beliefs about the correct answer are given by the posterior probabilities $p(a_i|s_k)$, which can be obtained from knowledge of the joint distribution of signals and answers. Beliefs about signals received by other respondents, say the probability of another respondent receiving signal s_j written as $p(s_j|s_k)$, are derived by computing the distribution of signals $p(s_j|a_i)$ conditional on a particular answer being correct, and marginalizing over all possible answers,

$$p(s_j|s_k) = \sum_i p(s_j|a_i)p(a_i|s_k)p(a_i)$$

More explicitly, one would write, $p(s_j^q|s_k^r) = \Pr(S^q = s_j|S^r = s_k)$, which is the probability that another, randomly selected respondent q receives signal s_j given that respondent r has received signal s_k. We omit the supercripts because the probability is the same for any pair of different respondents q, r.

As discussed in the main text and proven in Theorem 1 below, the probabilities $p(a_i|s_k)$ are always inconclusive, in that even an infinite sample of perfectly computed posterior probabilities over answers is compatible with any given answer being correct in some possible world model. Posterior probabilities strongly constrain the set of models with which they are compatible, but they do not identify the actual world.

Theorem 1. *The correct answer cannot be deduced by any algorithm relying exclusively on knowledge of actual signal probabilities, $p(s_k|a_{i^*}), k = 1, \ldots, n$ and posterior probabilities over answers implied by these signals, $p(a_i|s_k), k = 1, \ldots, n, i = 1, \ldots, m$.*

Proof. The proof is by construction of a possible world model that generates these signal probabilities and posterior probabilities for an arbitrarily selected answer.

Assume that the distribution of signals, $p(s_k|a_{i^*})$, and posterior probabilities, $p(a_j|s_k)$, are known but the correct answer a_{i^*} is unknown. We choose any answer a_i, and construct a corresponding possible world model $q(s_k, a_j)$ such q would generate the known signal distribution and posteriors if $i^* = i$.

Observe first that the known parameters do do not constrain the prior over signals, which we can set equal to:

$$q(s_k) = \frac{p(s_k|a_{i^*})}{p(a_i|s_k)}(\sum_j \frac{p(s_j|a_{i^*})}{p(a_i|s_j)})^{-1},\ k = 1, ..., n$$

Because posteriors must match observed posteriors: $q(a_j|s_k) = p(a_j|s_k)$, for $k = 1, \ldots, n$, $j = 1, \ldots, m$, the possible world model is now fixed: $q(s_k, a_j) = q(a_j|s_k)q(s_k)$. In particular, the prior over answers may be computed from the joint distribution,

$$q(a_i, s_k) = q(a_i|s_k)q(s_k) = p(s_k|a_{i^*})(\sum_j \frac{p(s_j|a_{i^*})}{p(a_i|s_j)})^{-1}$$

by summing over k:

$$q(a_i) = (\sum_j \frac{p(s_j|a_{i^*})}{p(a_i|s_j)})^{-1}$$

The marginal distributions $q(s_k)$, $q(a_i)$, together with the matching posteriors, $q(a_j|s_k) = p(a_j|s_k)$, for $k = 1, \ldots, n$, imply that if the correct answer is a_i, one would observe signal distribution $p(s_k|a_{i^*})$:

$$q(s_k|a_i) = \frac{q(a_i|s_k)q(s_k)}{q(a_i)} = p(s_k|a_{i^*})$$

Because a_i was freely chosen, this proves the theorem. □

Theorem 1 shows that the distribution of posterior probabilities over answers does not rule out any possible answer as the answer responsible for generating that distribution.

We turn therefore to the second type of beliefs, about signals received by other respondents. Because votes are functions of signals, ideal respondents receiving signal

s_k can compute the conditional probability $p(v_i|s_k)$ that another respondent will vote for a_i. For example, if the voting function instructs respondents to vote for the most likely answer, $V(s_j) = \arg\max_i p(a_i|s_j)$, to predict the probability that another respondent votes for a_i the respondent receiving signal s_k would add the probabilities of all signals j that are the most favorable to a_i:

$$p(v_i|s_k) = \sum_{j:V(s_j)=v_i} p(s_j|s_k) = \sum_{i=\arg\max_k p(a_k|s_j)} p(s_j|s_k)$$

Again, this notation suppresses respondent identity. In explict random variable notation, we would write $p(v_i|s_k)$, as $p(V^q = v_i|S^r = s_k)$ for $q \neq r$, i.e. the probability that an arbitrary respondent q votes for v_i. This is not to be confused with $p(V^r = v_i|S^r = s_k)$ which corresponds to stochastic voting by respondent r.

Similarly, we can define the joint distribution of votes (of an arbitrary respondent) and answers:

$$p(v_i, a_k) = \sum_{j:V(s_j)=v_i} p(s_j, a_j)$$

The conditional distributions, $p(v_i|a_k)$, and $p(a_k|v_i)$, are likewise well defined for any voting function.

2.3.2 The two worlds, many signals case $m = 2$, $n \geq 2$

We consider a more general version of the voting rule above, which allows us to avoid unanimity even when both signals favor the same answer. Specifically, we consider a cutoff based voting rule that instructs respondents to vote for a_1 if the probability of a_1 exceeds probability c_1, and for a_2 if the probability of a_2 exceeds $c_2 = 1 - c_1$. Formally, we can express this as

$$V(s_k) = \arg\max_i c_i^{-1} p(a_i|s_k)$$

The above voting rule is identical to the decision algorithm for an ideal observer in signal detection theory. If $c_1 = c_2 = 0.5$, the respondent is assumed to vote for the

more likely answer.

Theorem 2. *Assume that not everyone votes for the correct answer. Then the average estimate of the votes for the correct answer will be underestimated.*

Proof. We first show that actual votes for the correct answer exceed counterfactual votes for the correct answer, $p(v_{i^*}|a_{i^*}) > p(v_{i^*}|a_k), k \neq i^*$, as:

$$\frac{p(v_{i^*}|a_{i^*})}{p(v_{i^*}|a_k)} = \frac{p(a_{i^*}|v_{i^*})p(a_k)}{p(a_k|v_{i^*})p(a_{i^*})} = \frac{p(a_{i^*}|v_{i^*})}{(1-p(a_{i^*}|v_{i^*}))} \frac{(1-p(a_{i^*}))}{p(a_{i^*})}$$

The fraction on the right is well defined as $0 < p(a_{i^*}|v_{i^*}) < 1$; it is greater than one if and only if $p(a_{i^*}|v_{i^*}) > p(a_{i^*}|v_{i^*})p(v_{i^*}) + p(a_{i^*}|v_k)p(v_k) = p(a_{i^*})$, as $p(a_{i^*}|v_{i^*}) > c_{i^*}$, $p(a_{i^*}|v_k) < c_{i^*}$ by definition of the criterion based voting function.

A respondent with signal s_j computes expected votes by marginalizing across the two possible worlds, $p(v_{i^*}|s_j) = p(v_{i^*}|a_{i^*})p(a_{i^*}|s_j) + p(v_{i^*}|a_k)p(a_k|s_j)$. The actual vote for the correct answer is no less than the counterfactual vote, $p(v_{i^*}|a_{i^*}) \geq p(v_{i^*}|a_k)$. Therefore, $p(v_{i^*}|s_j) \leq p(v_{i^*}|a_{i^*})$, with strict inequality unless $p(a_{i^*}|s_j) = 1$. Because weak inequality holds for all signals, and is strict for some, the average predicted vote will be strictly underestimated. □

If there are more than two possible answers $m > 2$, the actual proportion of votes for the correct answer exceeds predictions provided that votes are defined by a cutoff vector $\sum_i c_i = 1$. However, it no longer points to a unique correct answer, as more than one answer may be underestimated.

2.3.3 Applying the surprisingly popular algorithm to binary questions

Theorem 2 shows that the average prediction of votes for the correct answer will underestimate the actual frequency of votes for the correct answer, or, in other words, the correct answer will be more popular than predicted. Applying this theorem to actual data is straightforward. Two quantities are elicited from each respondent: (1) which of the two options they personally vote for, (2) what fraction of the sample

they predict will vote for each option. The fraction of respondents voting for each option gives the actual vote frequency. The arithmetic mean of the predictions of all respondents gives the average vote prediction for each option. The surprisingly popular algorithm selects the answer that has an actual vote frequency that exceeds its predicted vote frequency.

2.3.4 The case $m = n > 2$

Our results for the general case with more than two answers and the same number of signals and answers, rely on a Lemma that shows how the ratio of posterior probabilities on the correct answer relative to any other answer can be derived from the signal frequencies, and their pairwise conditional probabilities. The Lemma is important because it expresses terms whose estimate requires knowing the correct answer (posterior probabilities on truth) as functions of terms that do not require knowledge of the correct answer.

Lemma. *Consider a possible world model with m answers and n signals and joint probability distribution $p(s_j, a_i)$. Let a_{i^*} denote the correct answer. Then:*

$$p(a_{i^*}|s_k) \propto p(s_k|a_{i^*}) \sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)}$$

(setting $0/0 \equiv 0$).

Proof. From Bayes' rule, we have,

$$p(s_i) = p(s_k)\frac{p(s_i|s_k)}{p(s_k|s_i)}$$

After summing over i, with $\sum_i p(s_i) = 1$, we solve for the prior probability of signal s_k:

$$p(s_k) = (\sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)})^{-1}$$

Invoking Bayes' rule again,

$$p(a_{i^*}|s_k) = \frac{p(s_k|a_{i^*})}{p(s_k)}p(a_{i^*}) = p(s_k|a_{i^*})\sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)}p(a_{i^*})$$

Because $p(a_{i^*})$ is constant across all k, the Lemma follows. □

The Lemma shows how the distribution of signals, and the pairwise predictions of signals, can identify the answer given by respondents who are best informed, in the sense of assigning the highest probability on the correct answer. These respondents would be least surprised by the correct answer, were it revealed.

To convert this Lemma into an algorithm for selecting the correct answer we need to assume that for each answer there is a unique signal such that respondents with that signal assign most probability to this answer, which is also more than the probability assigned to it by other respondents. This assumption is violated, for example, with the posteriors below:

$$p(a_i|s_k) = \begin{pmatrix} .4 & .3 & .3 \\ .45 & .55 & 0 \\ .2 & .3 & .5 \end{pmatrix}$$

If the correct answer is a_1 (first column), then respondents with s_2 (second row) would be least surprised, yet they would believe that the most likely correct answer is a_2. A selection principle based on treating as correct the answer selected by these respondents would incorrectly choose a_2. The theorem below rules out this possibility, by requiring that respondents voting for a given answer assign assign more probability to it than do respondents voting for other answers.

Theorem 3. *Assume $m = n$, $V(s_i) = v_i$, and $p(a_i|s_i) > p(a_i|s_j)$. Let a_{i^*} denote the correct answer. Define the prediction-normalized vote for a_k, $\overline{V}(k)$, as*

$$\overline{V}(k) = p(v_k|a_{i^*})\sum_i \frac{p(v_i|s_k)}{p(v_k|s_i)}$$

(setting $0/0 \equiv 0$). Then the correct answer has the highest prediction-normalized

votes.

Proof. Applying the Lemma, we have,

$$p(a_{i^*}|s_k) \propto p(s_k|a_{i^*}) \sum_i \frac{p(s_i|s_k)}{p(s_k|s_i)}$$

Because $V(s_i) = v_i$, we can rewrite this as:

$$p(a_{i^*}|s_k) \propto p(v_k|a_{i^*}) \sum_i \frac{p(v_i|s_k)}{p(v_k|s_i)} = \overline{V}(k)$$

$p(a_{i^*}|s_{i^*}) > p(a_{i^*}|s_k)$ by the assumption that respondents who vote for a given answer (including a_{i^*}) assign greater posterior probability to it than respondents voting for any other answer a_k. Therefore $\overline{V}(i^*) > \overline{V}(k)$, proving that the correct answer a_{i^*} has the highest prediction-normalized vote. □

One could apply Theorem 3 to experimental data using the following estimation procedure. Because a_{i^*} matches the actual world, the frequency of votes for answer a_k provides an estimate for $p(v_k|a_{i^*})$, which is exact in the limit of an infinite number of respondents. The probabilities $p(v_k|s_i)$ are estimated by asking respondents to predict the frequency of votes v_k and then averaging the predictions of those who voted v_i.

2.3.5 The case $m, n \geq 2$

It is possible to extend our approach to the general setting of m worlds, and n signals, provided one also elicits respondents' posterior distribution over possible answers.

Here we simply indicate the main idea, and for convenience consider $m > 2$, $n = 2$. That is, we have many coins, each with exactly two sides. Each individual thus has a small amount of information bearing on the many possible answers.

The bias of coin i is given by the ratio on the left side of the Bayesian identity below,

$$\frac{p(s_1|a_i)}{p(s_2|a_i)} = \frac{p(a_i|s_1)}{p(a_i|s_2)} \frac{p(s_1|s_2)}{p(s_2|s_1)}, \quad i = 1, ..., n$$

The analyst does not know these true coin biases, but can estimate them from the terms on the right, which respondents provide as their posterior probabilities, $p(a_i|s_1)$, $p(a_i|s_2)$, and pairwise predictions, $p(s_1|s_2)$, $p(s_2|s_1)$. The data therefore can be used to assign the correct bias to each possible coin.

To find the actual coin, the analyst asks respondents to report their toss outcome. The frequencies of observed tosses will converge to $p(s_1|a_{i^*})$ and $p(s_2|a_{i^*})$. The actual coin is then revealed as the coin whose assigned bias matches the observed one:

$$i = i^* \iff \frac{p(a_i|s_1)}{p(a_i|s_2)} \frac{p(s_1|s_2)}{p(s_2|s_1)} = \frac{p(s_1|a_{i^*})}{p(s_2|a_{i^*})}$$

A concrete example illustrates this. Assume three coins, a priori equally likely: (A) 2 : 1 biased for Heads, (B) 2 : 1 biased for Tails, (C) unbiased. From Bayes' rule respondents derive posterior probabilities over A,B,C as $(\frac{4}{9}, \frac{2}{9}, \frac{1}{3})$ following Heads, and $(\frac{2}{9}, \frac{4}{9}, \frac{1}{3})$ following tails. By symmetry of the assumptions, their predictions are also symmetric $p(s_j|s_k) = p(s_j|s_k)$.

Let us assume the actual coin is C. Respondents report their toss, their posterior probabilities on A, B, C, and their predicted toss distribution. From the predictions and posteriors the analyst assigns correct biases to the three possible coins, and notes that coin C is unbiased. Because toss reports converge to an even split between Heads and Tails, he deduces that the actual coin must be C.

The same method works in the general case, with more than two signals. It is important, however, that the elicitation separates signals (e.g., Heads vs. Tails) and possible states of the world (e.g., A, B, C). Respondents report signals, predict signals, and assign posteriors to states of the world.

2.4 Empirical tests of aggregation algorithms

2.4.1 Study descriptions

To test selecting the surprisingly popular answer (SP), we conducted studies with four types of semantic and perceptual content (details in Section 2.9). Studies 1a,b,c

used 50 U.S. state capitals questions, repeating the format (P) with different populations. Study 2 employed 80 general knowledge questions. Study 3 asked professional dermatologists to diagnose 80 skin lesion images as benign or malignant. Studies 4a,b presented 90 20th century artworks to laypeople and art professionals, and asked them to predict the correct market price category. All studies included a dichotomous voting question, yielding 490 items in total. Studies 1c, 2, and 3 additionally measured confidence. Predicted vote frequencies were computed by averaging all respondents' predictions.

2.4.2 Classification accuracy of aggregation methods

We first test pairwise accuracies of four algorithms: majority vote, SP, confidence weighted vote, and max-confidence, which selects the answer endorsed with highest average confidence. Across all 490 items, the SP algorithm reduced errors by 21.3% relative to simple majority vote ($p < 0.0005$ by two-sided matched-pair sign test). Across the 210 items on which confidence was measured, the reduction was 30.8% relative to majority vote ($p < 0.001$), 21.1% relative to confidence weighted vote ($p = 0.0107$), and 22.2% relative to max-confidence ($p = 0.0722$).

When frequencies of different correct answers are imbalanced, percentage agreement can be high by chance. Therefore we assess classification accuracy by categorical correlation coefficients, such as Cohen's kappa, F1-score, or Matthew's correlation. The SP algorithm has the highest kappa in every study (Figure 2-3); other coefficients yield similar rankings (Figures 2-4, 2-5, 2-6).

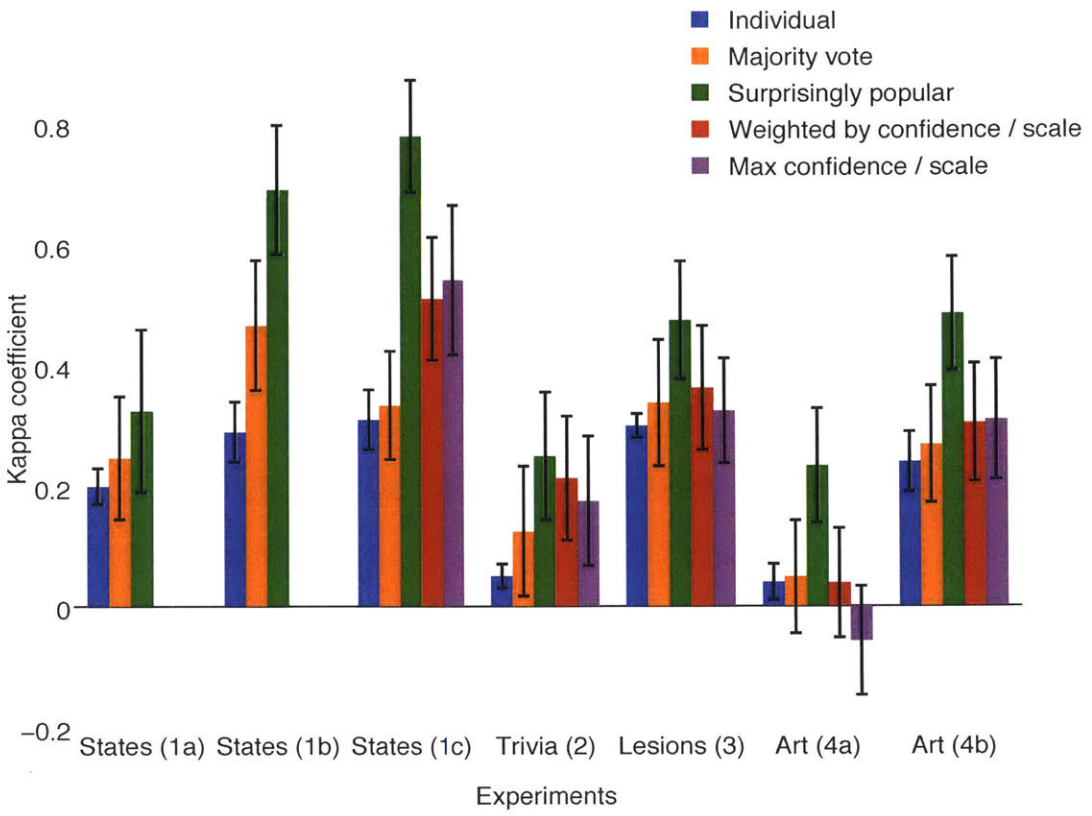


Figure 2-3: Results of aggregation algorithms on studies discussed in the text. N (items per study) = 50 (Studies 1abc), $N = 80$ (Studies 2 and 3), $N = 90$ (Studies 4ab). Agreement with truth is measured by Cohen's kappa, with error bars showing standard errors. kappa = $(A - B)/(1 - B)$, where A is percent correct decisions across items in a study, and B the probability of a chance correct decision, computed -according to answer percentages generated by the algorithm. Confidence was not elicited in Studies 1ab, 4ab. However, in 4ab we use scale values as proxy for confidence (Lebreton et al., 2015), giving extreme categories (on a four point scale) twice as much weight in scale-weighted voting, and 100% weight in max-scale. The results for Individual are average kappa-s across all individuals. SP is consistently the best performer across all studies. Results using Matthews correlation coefficient, F1-score, and percent correct are similar (Figs. 2-4, 2-5,2-6).

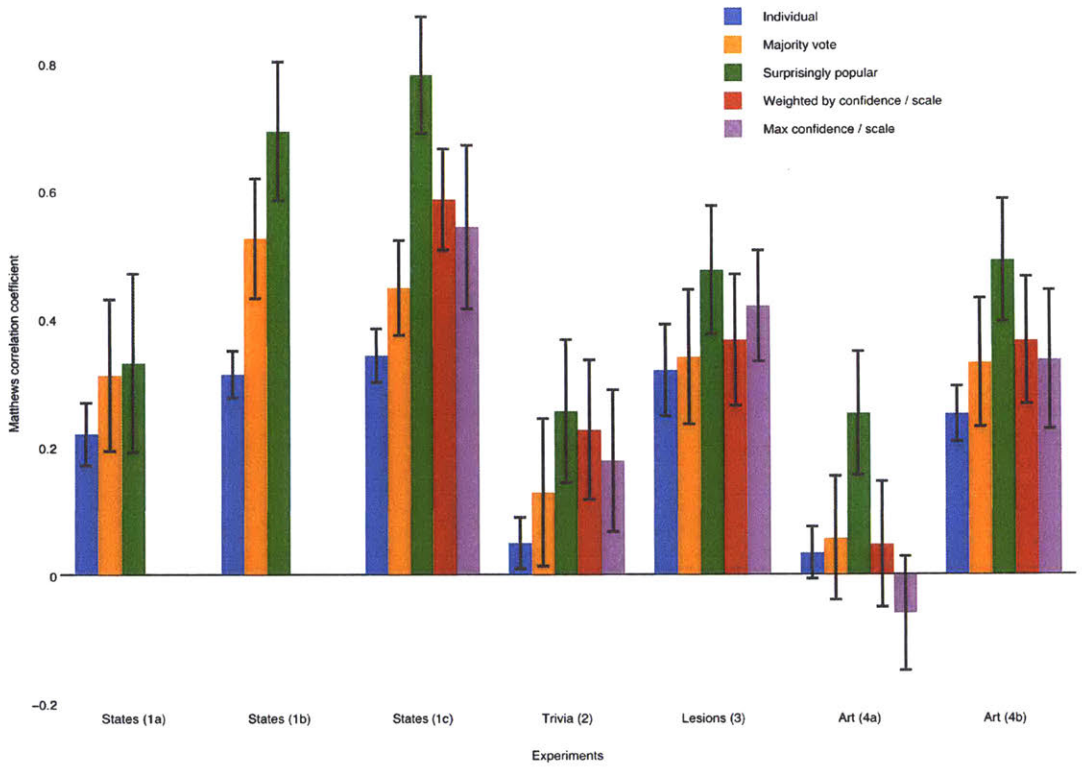


Figure 2-4: Performance of all methods across all studies, shown with respect to the Matthews correlation coefficient. Error bars are bootstrapped standard errors. Details of studies are given in Figure 2-3.

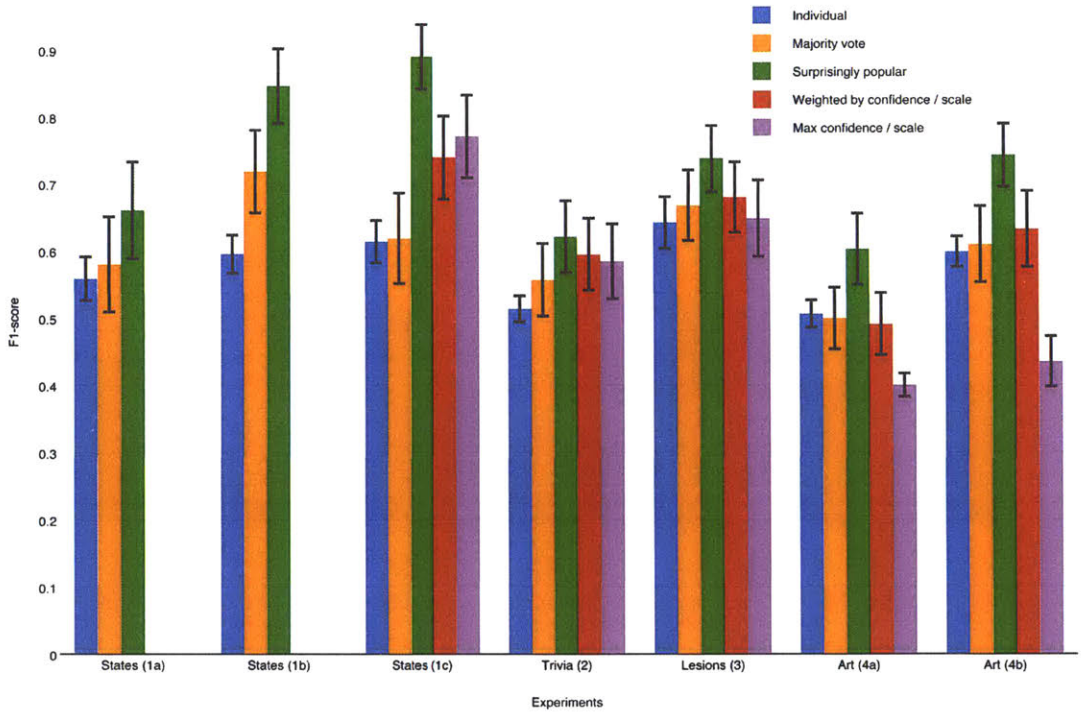


Figure 2-5: Performance of all methods across all studies, shown with respect to the macro-averaged F1-score. Error bars are bootstrapped standard errors. Details of studies are given in Figure 2-3.

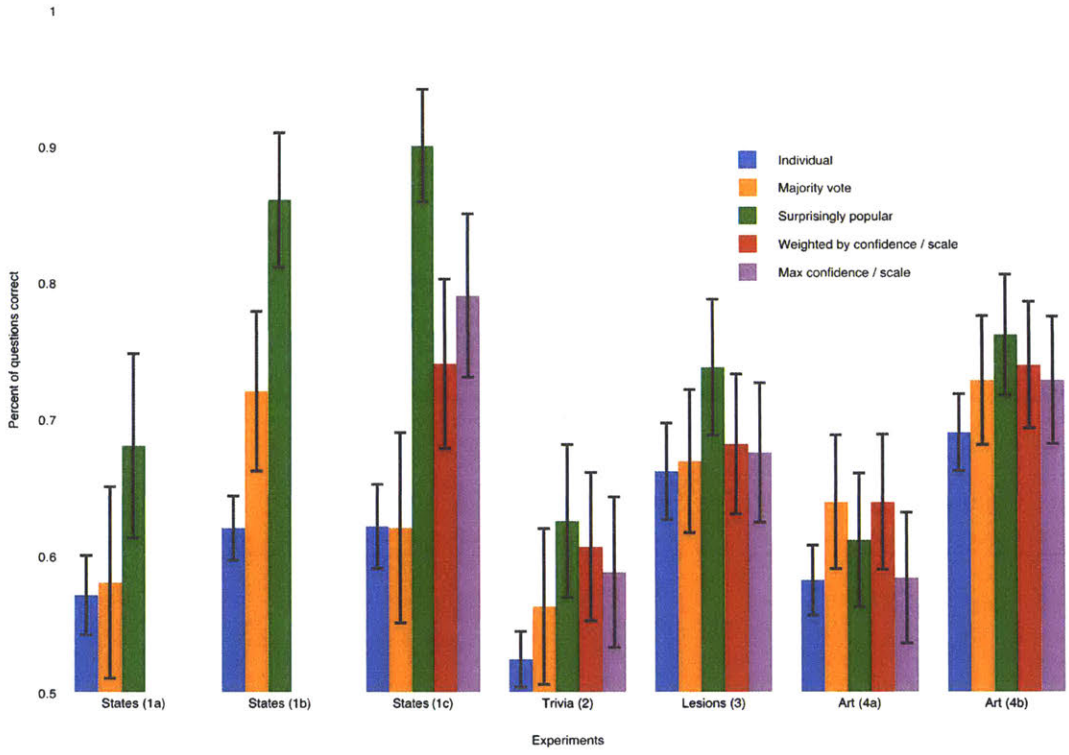


Figure 2-6: Performance of all methods across all studies, shown with respect to percentage of questions correct. Error bars are bootstrapped standard errors. Details of studies are given in Figure 2-3.

2.4.3 Art study results - preventing unwise crowds

The art domain, for which majority opinion is too conservative, provides insight into how SP works. Art professionals and laypeople estimated the price of 90 artworks by selecting one of four bins: $< \$1,000$; $\$1,000 - \$30,000$; $\$30,000 - \$1,000,000$; and $> \$1,000,000$. Respondents also predicted the binary division of their sample's votes relative to $30,000.

Both professionals and laypeople strongly favored the lower two bins, with professionals better able to discriminate value (Figure 2-7). The preference for low price is not necessarily an error: Asked to price an unfamiliar artwork, individuals may rely on their beliefs about market prices, and assume that expensive ($> \$30,000$) pieces are rare. This shared knowledge creates a bias when votes are counted, because similar, hence redundant, base rate information is factored in repeatedly, once for each respondent. Indeed, Figure 2-7 shows that the majority verdict is strongly biased against the high category. For example, facing a $100K artwork, the average professional has a 30% chance of making the correct call, while the majority vote of the professional panel is directionally correct only 10% of the time. It is difficult for any expensive artwork to be recognized as such by a majority. The SP algorithm corrects this by reducing the threshold of votes required for a high verdict, from 50% to about 25%.

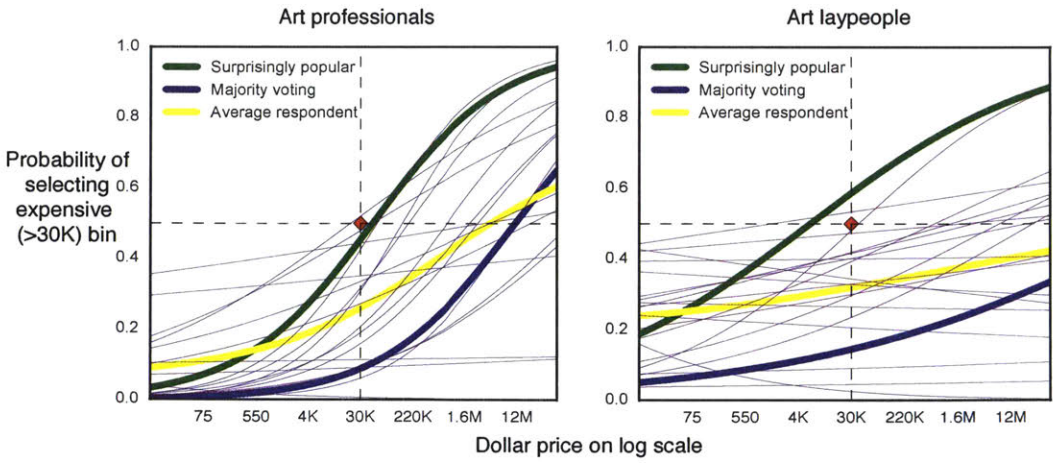


Figure 2-7: Logistic regressions showing the probability that an artwork is judged expensive (above $30K) as function of log actual market price. Thin purple lines are individual respondents in the art professionals and laypeople samples, and the yellow line shows the average respondent. Price discrimination is given by the slope of the logistic lines, which is significantly different from zero for 14/20 respondents in the professional sample, and 5/20 respondents in the laypeople sample (Chi-squared, $p < 0.05$). Performance is unbiased if a line passes through the red diamond, indicating that an artwork with true value of exactly $30K has a 50-50 chance of being judged above or below $30K. The bias against the higher price category, which characterizes most individuals, is amplified when votes are aggregated into majority opinion (blue line). The surprisingly popular algorithm (green line) eliminates the bias, and matches the discrimination of the best individuals in each sample.

2.4.4 Results on propositional knowledge

The two studies on propositional knowledge yielded different results (Figure 2-3). On capitals (Studies 1abc), SP reduced the number of incorrect decisions by 48% on capitals relative to majority vote. SP was less effective on the knowledge questions in Study 2 (14% error reduction, $p = .031$, two-sided matched-pair sign test). This is the only study that used the Amazon Mechanical Turk respondent pool. In contrast to other studies, the predicted vote splits in Study 2 were within 10% of 50% for 81% of

items, compared to 22% of such items across other studies. This limited opportunities for SP to alter majority vote.

2.4.5 States capitals results - Consensus and systematic error

The simulation results in the next section predict that voting accuracies track consensus while the accuracy of surprisingly popular answers is independent of consensus. and so to illustrate this point we show the accuracy of three different aggregation methods as a function of consensus on Study 1a,b,c in Figure 2-8. Furthermore, because the states capitals studies were run three times we can use this to investigate the extent to which selecting the surprisingly popular answer and voting differ at sytematically obtaining the wrong answer for particular questions. (Confidences were only elicited in Study 1c, and so this analysis does not include confidence-weighted voting.) Figure 2-8 includes histograms for voting and selecting the surprisingly popular answer showing for how many of the fifty states the method was correct for the 0, 1, 2, or 3 times the study was run. As these histograms show, there are a number of states for which voting is systemtically wrong every time the study was run, but this is not the case for selecting the suprisingly popular answer - the surprisingly popular answer gets a few states incorrect each time the study is run, but it is not the same set of states each time.

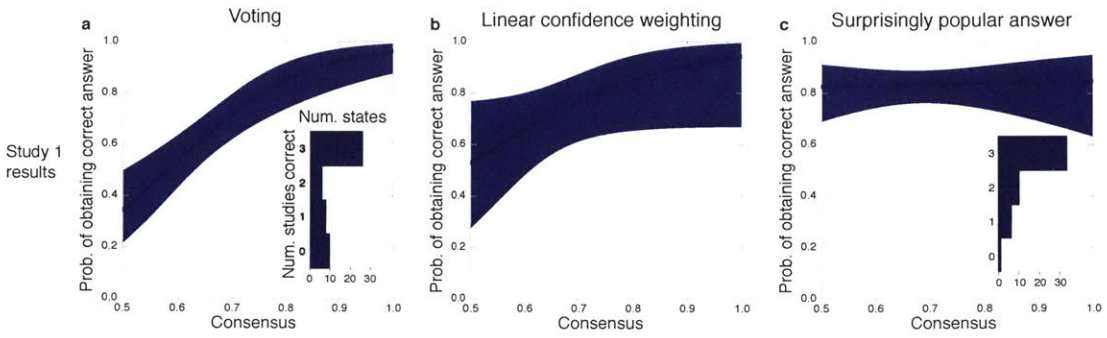


Figure 2-8: Logistic curves (with 95% confidence intervals) show the accuracy of the methods as a function of consensus, treating the Study 1a,b,c results described in the text as providing 150 questions. Histograms show the number of states for which the answer was correctly determined (using voting and the surprisingly popular answer) for none, one, two, or all of the experiments. Confidence-weighted voting could only be applied to one study, and so does not include a histogram.

2.5 Simulations for finite samples

Our theoretical results that justify selecting the surprisingly popular answer all depend on the assumption of an infinite sample of Bayesian respondents. To investigate the performance of our new method, as well as standard aggregation mechanisms, on finite samples of Bayesian respondents we turn to numerical simulation.

2.5.1 Sampling assumptions

The simulations were based on the biased coin model (Figure 2-2). The world prior, coin biases, the actual world, and respondent coin flips were randomly generated to produce simulated finite samples of votes, confidences, and vote predictions (Figure 2-9).

We performed numerical simulations for the simplest $m = n = 2$ case, under a uniform sampling assumption. We randomly sampled 1000 datasets each with 50 questions, answered by up to 1000 respondents. For each dataset, respondent subsets

of size $N \in \{6, 12, 33, 1000\}$ were randomly sampled. For each question $w = 1, 2, \ldots$ a possible world model consisting of a joint distribution of two world states and two signals $p(s_k^w, a_i^w)$ was uniformly sampled, with resampling if it did not satisfy $p(a_k^w | s_k^w) > 0.5$. An actual world $a_{i^*}^w$ was sampled given $p(a_i^w)$, and signals were sampled given $p(s_k^w | a_{i^*}^w)$. Votes, confidences, and vote predictions were computed for each ideal Bayesian respondent.

2.5.2 Simulation results

Under these sampling assumptions, individuals are correct 75% of the time. Applying majority voting gives an accuracy of 86%. This 11% improvement is the standard wisdom-of-the-crowd effect. SP is almost infallible for large samples, and it shows good, though not perfect, performance even on small sample sizes. However, given the 86% accuracy of majority vote, SP may need many problems to demonstrate a statistically significant advantage. For example, with 50 problems and $n = 30$, the SP superiority attains $p < .05$ for only 40% of simulated studies.[1]

2.6 Analysis of predictions of the votes of others

Our model describes how Bayesian respondents formulate a prediction of the vote distribution of others based on their received signal. Here, we describe some descriptive statistics of the predictions of votes. Respondents' votes were, in general, correlated with their own answers, which is one reason that vote predictions tend not to be simply 50%-50% . That is, respondents voting for option A, compared to those voting for option B, put higher probability on other respondents also voting for option A. In other words, respondents demonstrated self-projection or false consensus (Ross et al.,

[1]We sampled 1000 synthetic datasets each consisting of 50 questions and answered by 30 respondents. We computed the answer given by majority vote and the surprisingly popular answer and for each dataset performed a paired-samples t-test of voting correctness against the correctness of the surprisingly popular answer. For 392 of the sampled datasets, the test showed a significant advantage ($p < 0.05$) of the surprisingly popular answer over majority voting, and for none of the sampled datasets was there a significant advantage of majority voting over the surprisingly popular answer.

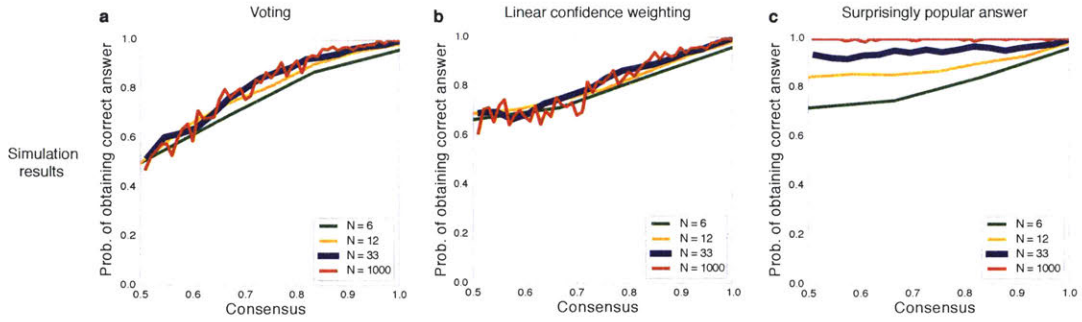


Figure 2-9: Performance of aggregation methods on simulated datasets of binary questions, under uniform sampling assumptions. A pair of coin biases (i.e. signal distribution parameters), and a prior over worlds are sampled, each from independent uniform distributions. Combinations of coin biases and prior that result in recipients of both coin tosses voting for the same answer are discarded. An actual coin is sampled according to the prior, and tossed a finite number of times to produce the votes, confidences, and vote predictions required by the different methods (see Section 2.5.1 for simulation details). As well as showing how sample size affects different aggregation methods the simulations also show that majorities become more reliable as consensus increases. A majority of 90% is correct about 90% of the time, while a majority of 55% is not much better than chance. This is not due to sampling error, but reflects the structure of the model and simulation assumptions. According to the model, an answer with x% endorsements is incorrect if counterfactual endorsements for that answer exceed x% (Theorem 2), and the chance of sampling such a problem diminishes with x.

1977). As Dawes originally argued (Dawes, 1989, 1990), such a false-consensus effect need not be irrational and can be entirely consistent with Bayesian reasoning. The idea is that respondents own beliefs act as an 'informative sample of one' and thus give them information about the distribution of beliefs in the population. Hence, not only is the existence of the false-consensus effect not an argument against selecting the surprisingly popular answer, it is actually predicted by the possible world model. Across the three capitals studies (1a,b,c), respondents voting for False predicted that 49% of respondents would endorse True, whereas those voting for True predicted that 73% would. For the trivia questions in Study 2, this prediction was 45% and 59%, respectively. In the lesion study, respondents who voted for benign predicted that 34% of respondents would vote for malignant, but those voting for malignant predicted 79%. In the art study with MIT students, those voting for the expensive

price bin predicted that 46% of other would, whereas those voting for cheap predicted that only 22% would vote for expensive. For the study with art professionals, these predictions were 53% and 14%, respectively.

For each study, we can examine how often respondents gave predictions that were close to a uniform split. For studies where predictions were given as percentages, we count the fraction of times that a vote prediction is given that is within 10% of 50-50, i.e. the prediction of votes for the first option is between 40% and 60%. Across the three states studies, an average of 36% of predictions were within 10% of 50-50, in the trivia study 56%, in the art MIT study 29%, and in the art professionals study 19%. For the lesions study, respondents gave their predictions on an 11 point scale, and 30% of predictions were one of the three middle bins. Note that if all respondents simply predict that 50% of the sample will endorse each of two possible answers, then the surprisingly popular answer is the same as that obtained by majority vote.

As a further description of the predictions that people made, we compare predictions of the perecentage of people voting True (or malignant) to the probability that people put on the answer being True (or malignant), inferred from their vote and confidence. For the MIT states study where confidence was elected this correlation is $r_S = 0.64$, ($p < 0.001$), for the lesions study $r_s = 0.87$, ($p < 0.001$), and for the trivia study $r_S = 0.48$, ($p < 0.001$).

While false consensus or social projection has been extensively studied in social psychology, the phenomena of false-uniqueness whereby respondents insufficiently weight their own answers is much less studied (Chambers, 2008; Suls and Wan, 1987). There are a number of approaches to detect false-uniqueness. We regress for each respondent individually, across all of the questions that they answer, the actual fraction of people endorsing a given answer for each question against the respondent's prediction of the fraction of people endorsing that answer together with the respondent's own answer. We can then test, for each respondent, whether the coefficient on their own answer is significant. That is, we can assess whether there is information in a respondent's own answer about the fraction of people endorsing an answer, that is not captured in the respondent's predictions. As a result of this analysis, we find

that there are indeed respondents participating in our studies that exhibit such false-uniqueness, in the range of 50% to 75% of respondents depending on the study. Unlike the false-consensus effect, false-uniqueness is not predicted by a model of Bayesian respondents, but the argument for selecting the surprisingly popular answer is robust to the existence of respondents who display false-uniqueness since even if a respondent combines their prior and likelihood but puts too much weight on their prior belief about which world is likely and too little weight on their signal, their prediction will still fall between the belief distribution in each counterfactual world.

We do not have sufficient experimental evidence to justify a particular method of eliciting useful vote predictions, but we offer a few suggestions, and speculations for future testing. Vote predictions can be incentivized for accuracy. For example, respondents can be paid a bonus which depends on the Kullback-Leibler divergence between their prediction and the actual distribution, or incentivized more generally using the Bayesian Truth Serum (Prelec, 2004). Respondents can be explicitly encouraged to consider whether they are in the minority or majority, and what opinions people different to themselves may hold. Instructions may help respondents recognize cases where despite them having high confidence in an answer, they should also believe that only a minority of respondents would vote for this answer. It is possible that choosing to not elicit confidence prior to eliciting predictions may help respondents to avoid conflating these two quantities. When dealing with respondents answering multiple questions who give identical vote predictions for every question, one could take steps to encourage them to reflect on whether this is an accurate reflection of their beliefs. Respondents could be given information about the composition of the sample that they are in, to aid them in making good predictions about the answers of others.

2.7 Discussion

SP performance will always be limited by the information available to the respondent sample, and the competence of respondents. If the available evidence is incomplete

or misleading, the answer that best fits the evidence may be incorrect. This qualifier can be made explicit by careful phrasing of questions. A question like "Will global temperature increase by more than 5%?" could be worded as: "Given current evidence, is it more likely or not that global temperature will increase by more than 5%?"

The SP algorithm is robust to some deviations from ideal responding. The outcome will not change, for example, if respondents simply predict the vote frequency consistent with the world they believe most likely, instead of computing confidences and interpolating predictions. Similarly, the surprisingly popular answer is robust to respondents who give a prediction obtained from mixing the vote frequencies in each world according to the prior over worlds rather than their posterior distribution over worlds. Even if some respondents find the prediction task too difficult, they may simply predict the default value of 50-50. This would bring SP results closer to majority opinion.

While selecting the surprisingly popular answer is robust to some of the biases that people's predictions may exhibit, such as false-uniqueness effects, there are other psychological biases that could decrease the accuracy of selecting the surprisingly popular answer if they were prevalent in a sample of respondents. In particular, if respondents who gave the correct answer were in a minority but exhibited strong curse of knowledge effects (Nickerson, 2001; Keysar et al., 1995; Nickerson, 1999; Hinds, 1999; Kennedy, 1995; Camerer et al., 1989) or 'truly false consensus effects' (Krueger and Clement, 1994) and incorrectly believed that most other people also knew the correct answer, then the surprisingly popular answer could be incorrect.

A different kind of objection to selecting the surprisingly popular answer one might call the 'crazy, secret sect objection'. Suppose that there is a secret sect whose leader has told the members that the world will end tomorrow. If everyone in the world was then surveyed about whether the world would end tomorrow, would selecting the surprisingly popular answer incorrectly conclude that the world would end tomorrow?[2] It is possible that people may have sensible beliefs about the prevalence

[2]Since the members of the sect will vote that the world will end tomorrow and this would not be

of crazy sects in that they think there is always some small number, but believe that there are more when the world is actually ending, but suppose they underestimate the frequency of such beliefs in this case. Which assumptions of our model have been violated? Sometimes selecting the surprisingly popular answer will select the answer given by a group of people with extra information, but such extra information may be incorrect, and point in the wrong direction. Alternatively, the problem may be that the possible world model makes the assumption that people share knowledge of the joint distribution of signals and states of the world, but this may be inaccurate, for example when there is a crazy sect receiving a signal that other people do not put the correct probability on being received or when people inside and outside the sect have a different prior probability over world states.

Ultimately, one would like to apply our method to potentially controversial problems, like political and environmental forecasts, where it is important to guard against manipulation. For example, respondents might try to increase the chance that their favored option wins by strategically submitting low vote predictions for that option. When this is a concern, one can impose truth telling incentives using the Bayesian truth serum (John et al., 2012; Prelec, 2004). This mechanism also requires vote predictions and rewards respondents for answers that are revealed to be surprisingly popular. Here, we have shown that the same criterion selects the best collective answer.

The SP algorithm can be compared to a prediction market, which also aggregates opinions on single questions (Arrow et al., 2008). Both methods allow experts to override the majority view, and both associate expertise with choosing alternatives whose popularity exceeds current expectations. However, unlike prediction markets, our method accepts non-verifiable propositions, such as counterfactual conjectures in public policy, history or law. This, together with the simple input requirements, greatly expands its application range.

predicted by everyone else, implying that their answer is surprisingly popular.

2.8 Conclusion

Although democratic methods of opinion aggregation have been influential and productive, they have underestimated collective intelligence in one respect. People are not limited to stating their actual beliefs; they can also reason about beliefs that would arise under hypothetical scenarios. Such knowledge can be exploited to recover truth even when traditional voting methods fail. If respondents have enough evidence to establish the correct answer, then the surprisingly popular principle will yield that answer; more generally, it will produce the best answer in light of available evidence. These claims are theoretical and do not guarantee success in practice, as actual respondents will fall short of ideal. However, it would be hard to trust a method if it fails with ideal respondents on simple problems like (P). To our knowledge, the method proposed here is the only one that passes this test.

2.9 Appendix - Methods

2.9.1 Informed consent

All studies were approved by the M.I.T. Committee on the use of humans as experimental subjects (COUHES). For all studies, informed consent was obtained from respondents using text approved by COUHES. For in-person studies, respondents signed a consent form and for online studies, respondents checked a box.

2.9.2 Studies 1a, b – State capitals

Materials and methods

The survey instrument consisted of the following single sheet of paper (shown in Figure 2-10), which respondents were asked to complete. There was no time limit.

Your CODE _____

For each statement, please answer whether you think it is **True (T)** or **False (F)**
and then ***estimate the percentage of participants in this experiment that will answer "True".***

	Your Answer	% Answering "True"		Your Answer	% Answering "True"
Birmingham is the capital of Alabama.	____	_____ %	Billings is the capital of Montana	____	_____ %
Anchorage is the capital of Alaska.	____	_____ %	Omaha is the capital of Nebraska.	____	_____ %
Phoenix is the capital of Arizona.	____	_____ %	Las Vegas is the capital of Nevada.	____	_____ %
Little Rock is the capital of Arkansas.	____	_____ %	Manchester is the capital of New Hampshire.	____	_____ %
Los Angeles is the capital of California.	____	_____ %	Newark is the capital of New Jersey.	____	_____ %
Denver is the capital of Colorado.	____	_____ %	Albuquerque is the capital of New Mexico.	____	_____ %
Bridgeport is the capital of Connecticut.	____	_____ %	New York City is the capital of New York.	____	_____ %
Wilmington is the capital of Delaware.	____	_____ %	Charlotte is the capital of North Carolina.	____	_____ %
Jacksonville is the capital of Florida.	____	_____ %	Fargo is the capital of North Dakota.	____	_____ %
Atlanta is the capital of Georgia.	____	_____ %	Columbus is the capital of Ohio.	____	_____ %
Honolulu is the capital of Hawaii.	____	_____ %	Oklahoma City is the capital of Oklahoma.	____	_____ %
Boise is the capital of Idaho.	____	_____ %	Portland is the capital of Oregon.	____	_____ %
Chicago is the capital of Illinois.	____	_____ %	Philadelphia is the capital of Pennsylvania.	____	_____ %
Indianapolis is the capital of Indiana.	____	_____ %	Providence is the capital of Rhode Island.	____	_____ %
Des Moines is the capital of Iowa.	____	_____ %	Columbia is the capital of South Carolina.	____	_____ %
Wichita is the capital of Kansas.	____	_____ %	Sioux Falls is the capital of South Dakota.	____	_____ %
Lexington is the capital of Kentucky.	____	_____ %	Memphis is the capital of Tennessee.	____	_____ %
New Orleans is the capital of Louisiana.	____	_____ %	Houston is the capital of Texas.	____	_____ %
Portland is the capital of Maine.	____	_____ %	Salt Lake City is the capital of Utah.	____	_____ %
Baltimore is the capital of Maryland.	____	_____ %	Burlington is the capital of Vermont.	____	_____ %
Boston is the capital of Massachusetts.	____	_____ %	Virginia Beach is the capital of Virginia.	____	_____ %
Detroit is the capital of Michigan.	____	_____ %	Seattle is the capital of Washington.	____	_____ %
Minneapolis is the capital of Minnesota.	____	_____ %	Charleston is the capital of West Virginia.	____	_____ %
Jackson is the capital of Mississippi.	____	_____ %	Milwaukee is the capital of Wisconsin.	____	_____ %
Kansas City is the capital of Missouri.	____	_____ %	Cheyenne is the capital of Wyoming.	____	_____ %

Figure 2-10: Survey instrument for Study 1a,b

Respondents and procedure

Study 1a was conducted in the context of two MIT, Sloan MBA classes. 51 respondents were asked to mark their answer sheet by a personal code, and were promised feedback about the results, but no other compensation. Study 1b was conducted at the Princeton Laboratory for Experimental Social Science (PLESS, http://pless.princeton.edu/). 32 respondents were drawn from the pool of pre-registered volunteers in the PLESS database, which is restricted to Princeton students (undergraduate and graduate). Respondents received a flat $15 participation fee. In addition, the two respondents with the most accurate answers received a $15 bonus, as did the two respondents with the most accurate percentage predictions. (In fact, one respondent received both bonuses, earning $45 in total). Respondents marked their sheet by a pre-assigned code, known only to the PLESS administrator who distributed the fee and bonus.

2.9.3 Study 1c – State capitals

Materials and methods

The survey was administered on a computer, and a screenshot from the experiment is shown in Figure 2-11.

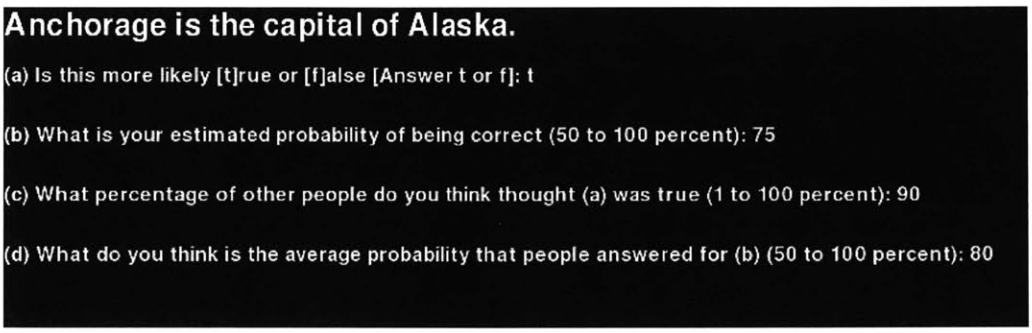


Figure 2-11: Screenshot of question from Study 1c.

Respondents and procedure

The study was conducted in the MIT Behavioral Research Lab (http://web.mit.edu/brl/). 33 respondents were recruited from the MIT Brain and Cognitive Sciences Department experimental respondents mailing list, with participation restricted to members of the MIT community. Respondents received a $15 participation fee. In addition, the top 20% of respondents with the most accurate answers with respect to ground truth and the top 20% of respondents with the most accurate predictions about the beliefs of others earned a $25 bonus. Respondents were eligible to receive both bonuses. The explanation given to respondents about the bonus system is reproduced below.

Determination of bonus:

After the study is complete, we will calculate two accuracy scores for each respondent.

(1). Your objective accuracy score is based on your answers to (a) and (b).

For each statement we calculate the probability that you think the statement is true and use this, together with whether the statement is actually true to calculate your score. We use the Brier scoring function, which is designed so that your score is maximized when you report your true guess and confidence level. Below is a table which helps you understand the score you would receive, depending on whether your answer in (a) was correct or incorrect. The table gives the score at intervals of ten percentage points, but you can choose any percentage between 50% and 100%.

Your confidence	*Score if (a) correct*	*Score if (a) incorrect*
50%	*0*	*0*
60%	*9*	*-11*
70%	*16*	*-24*
80%	*21*	*-39*
90%	*24*	*-56*
100%	*25*	*-75*

Points to note:

- *the more certain you claim to be, the more points you can win*
- *as you approach 100%, the penalty for being incorrect climbs much faster than the gains for being correct.*

A tip:

- *In the long run, you will score the most points if the numbers correspond to your true levels of confidence. Expressing too much confidence is a common mistake in this game.*

(2). Your prediction accuracy score is based on your answers to (c) and (d).

Your prediction accuracy score reflects how well you have predicted the actual percentages of respondents who answered Yes to each of the fifty questions, and how well you have estimated the average confidence levels.

2.9.4 Study 2 – General knowledge questions

Materials and Methods

The survey consisted of 80 trivia questions in the domains of history, language, science, and geography. The survey was administered as an online questionnaire and question order was randomized across respondents. The questions were a subset of the 150 questions from the True/False quizzes in these domains on the quiz site Sporcle (www.sporcle.com). Two online pilot experiments (of 70 and 80 questions each) were conducted in which respondents were only asked whether they thought the answer to each question was True or False, i.e respondents were not asked to make second-order predictions. Using the results of the two pilot experiments, 80 questions were selected by matching the questions for percentage correct, e.g. a question that 30% of respondents answered correctly was matched with a question that 70% of respondents answered correctly. This resulted in a balanced final survey with respect to the number of questions the majority answered correct as well as the number of questions for which the correct answer was false, as shown by the contingency table in Table 2.1.

	Actual answer is false	Actual answer is true	Totals
Majority incorrect	20	19	39
Majority tie	1	1	2
Majority correct	19	20	39
	40	40	80

Table 2.1: Contingency table showing distribution of questions for Study 2.

Example questions, together with the percentage of respondents who answered correctly in the pilot experiment are shown in Table 2.2.

Example question	Percent of respondents correct in pilot experiments
Japan has the world's highest life expectancy	10
The Nile River is more than double the length of the Volga	20
Portuguese is the official language of Mozambique	30
Avogadro's constant is greater than Planck's constant	40
The currency of Switzerland is the Euro	50
Abkhazia is a disputed territory in Georgia.	50
The chemical symbol for Tin is Sn	60
The Iron Age comes after the Bronze Age	70
Schuyler Colfax was Abraham Lincoln's Vice President	80
The longest bone in the human body is the femur	90

Table 2.2: Example questions from Study 2 and percent correct in pilot experiments.

Respondents were given the following instructions:

Please read the following 80 True/False trivia questions carefully and make your best guess.

For each question, we'll ask you to do three things:

(a) *Say whether you think the statement is more likely True or False*

(b) *Think about your own beliefs and estimate the probability that your answer is correct*

(c) *Think about other people's beliefs and predict the percentage of people who guessed the answer was 'True'*

To give an estimate of the probability that their answer was correct, respondents chose one of the six following options:

(a) Totally uncertain, a coin toss (about 50% chance of being correct)

(b) A little confident (about 60% chance of being correct)

(c) Somewhat confident (about 70% chance of being correct)

(d) High confidence (about 80% chance of being correct)

(e) Very high confidence (about 90% chance of being correct)

(f) Certain (about 100% chance of being correct)

To answer the second-order prediction, people gave a percentage.

Respondents were asked to not search for the answers to the questions. Respondents searching for the answer, rather than answering from their own knowledge, does not make affect testing the aggregation method since this is simply an additional source of information for some respondents who may thus be more accurate. The average time to complete all three parts of a question was 17 seconds and it was not the case that if a respondent took more time to answer a question they were more likely to be correct, suggesting that, in fact, searching for the correct answer was not common.

Respondents and Procedure

Respondents were recruited from Amazon Mechanical Turk and were paid a flat fee of $5.00 with 39 respondents completing the survey. Respondents who took part in either of the pilot experiments were excluded from participating in the final experiment.

2.9.5 Study 3 – Dermatologists assessing lesions

Materials and Methods

The survey was administered online. Respondents were divided into two groups, with one survey containing images of 40 benign and 20 malignant lesions, and the other survey containing images of 20 benign and 40 malignant lesions. The 80 images used in the experiment were obtained from Atlas Dermatologico, DermIS, and DermQuest. The images were selected to be approximately the same size, had no visible signs of

biopsy, and were filtered for quality by an expert dermatologist. Question order was randomized across respondents. Since all lesions pictured in the survey had been biopsied, whether a particular lesion was benign or malignant was known to us.

For each image of a lesion, respondents predicted whether the lesions was benign or malignant, gave their confidence on a six point Likert scale from 'absolutely uncertain' to 'absolutely certain' and estimated the likely distribution of opinions amongst other dermatologists on an eleven point scale from 'perfect agreement that it is benign' to 'perfect agreement that it is malignant' with the midpoint labeled as 'split in opinions with equal number of benign and malignant diagnoses'.

Respondents and procedure

Dermatologists were recruited by referral and 25 respondents answered the survey, with 12 in the condition with 40 benign lesions and 13 in the condition with 20 benign lesions. Respondents had an average of 10.5 years of experience. Respondents were told that a $25 donation would be made to support young investigators in dermatology for every completed survey, and that if the survey was completed by a particular date this would be increased to $50. Respondents were also told that a randomly selected respondent would receive $1000.

2.9.6 Study 4a, b – Professionals and laypeople judging art

Materials and Methods

The survey instrument consisted of a bound booklet with each page containing a color picture of a 20th century art piece and questions about the piece. The medium and dimensions were given for each piece. Participants were given the following instructions:

The survey contains 90 reproductions of modern (20th century) artworks. For each artwork we will aks you a few questions.

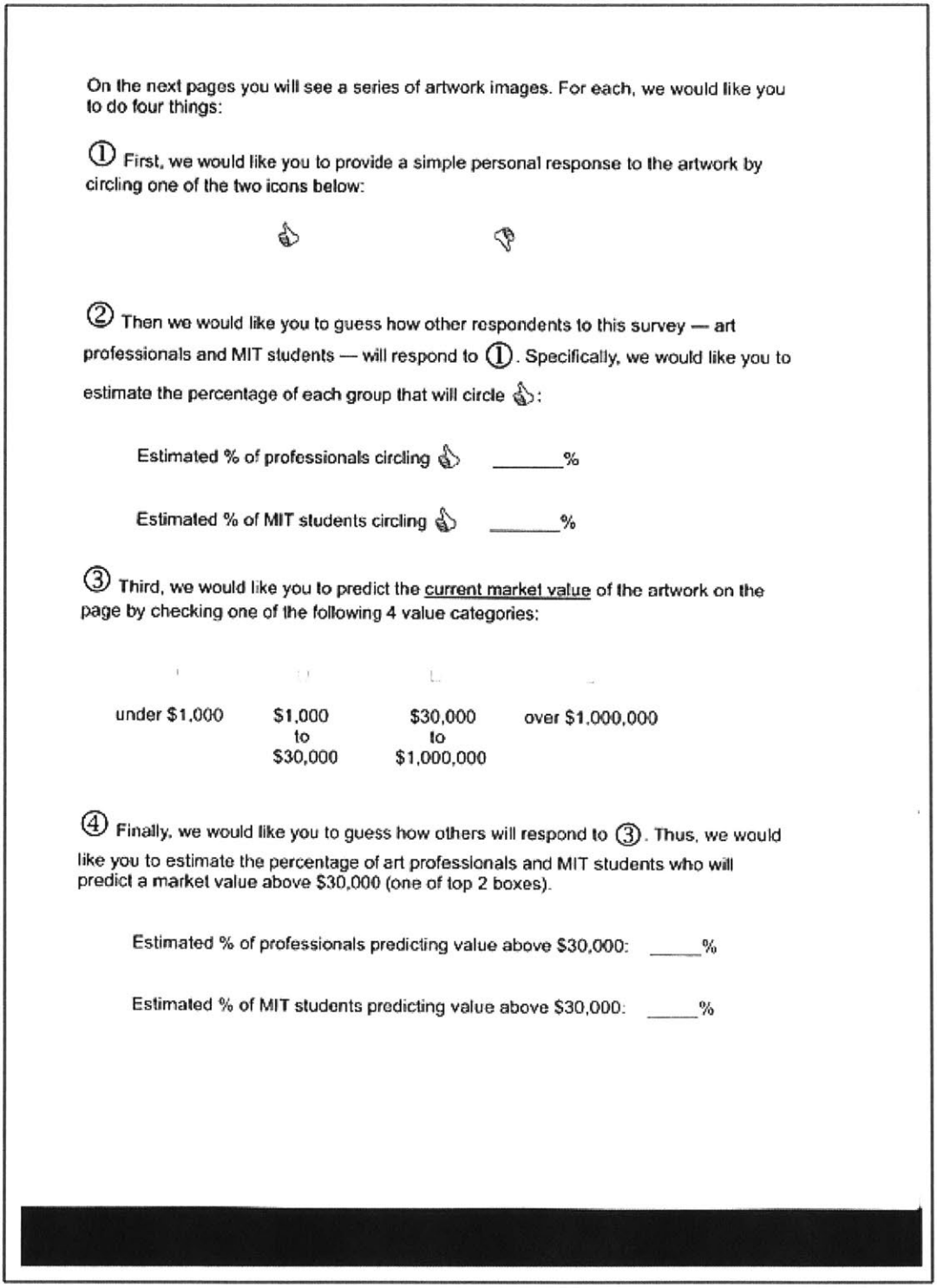
Your answers will help us understand how professionals and non-professionals respond to modern art.

- *By professionals, we have in mind people working with art, in galleries or museums.*
- *By non-professionals, we are referring to MIT master's and doctoral students who have not taken any formal art or art history classes.*

We are also interested in how well people can predict the responses of other people. So, some questions will ask you to guess how other people will respond.

This will be explained more fully on the next page. If there is anything unclear about our instructions please do not hesitate to ask!

We reproduce the page of the booklet containing the instructions and an example question page in Figures 2-12 and 2-13.


On the next pages you will see a series of artwork images. For each, we would like you to do four things:

① First, we would like you to provide a simple personal response to the artwork by circling one of the two icons below:

② Then we would like you to guess how other respondents to this survey — art professionals and MIT students — will respond to ①. Specifically, we would like you to estimate the percentage of each group that will circle :

Estimated % of professionals circling _____%

Estimated % of MIT students circling _____%

③ Third, we would like you to predict the current market value of the artwork on the page by checking one of the following 4 value categories:

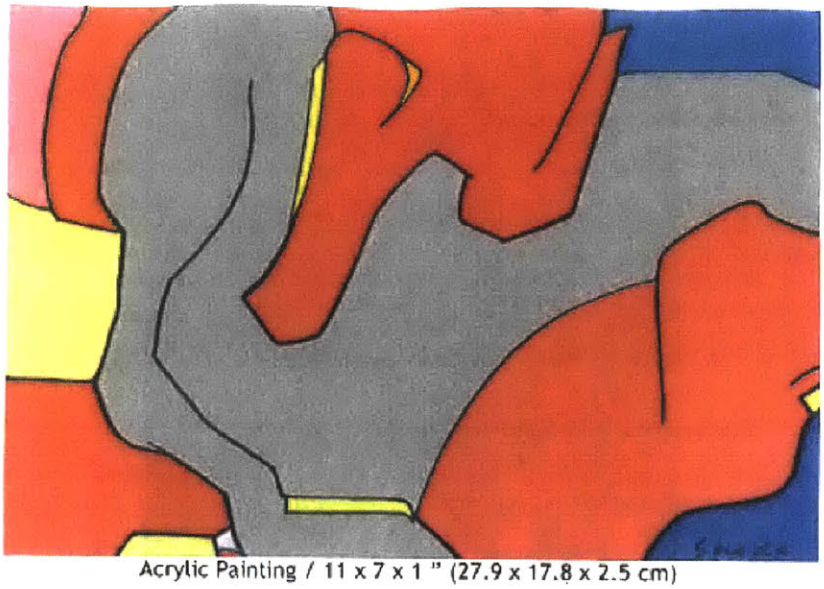
under $1,000	$1,000 to $30,000	$30,000 to $1,000,000	over $1,000,000

④ Finally, we would like you to guess how others will respond to ③. Thus, we would like you to estimate the percentage of art professionals and MIT students who will predict a market value above $30,000 (one of top 2 boxes).

Estimated % of professionals predicting value above $30,000: _____%

Estimated % of MIT students predicting value above $30,000: _____%

Figure 2-12: Page of booklet containing instructions for Study 4a,b.



Acrylic Painting / 11 x 7 x 1 " (27.9 x 17.8 x 2.5 cm)

① Your response:

② Estimated % of professionals circling _____%

Estimated % of MIT students circling _____%

③

under $1,000	$1,000 to $30,000	$30,000 to $1,000,000	over $1,000,000

④ Estimated % of professionals predicting value above $30,000: _____%

Estimated % of MIT students predicting value above $30,000: _____%

Figure 2-13: Example page of booklet given to respondents in Study 4.

Respondents and Procedure

Two groups of respondents completed the survey. The MIT group consisted of twenty MIT graduate students who had not taken courses in art or in art history. They were paid $20 as compensation for their time. Respondents came individually into the lab, and completed the survey in a room alone. The Newbury group, named for Newbury street in Boston which has many art galleries, consisted of art professionals – predominantly managers of art galleries. The art professionals were visited by appointment at their offices and completed the survey during the appointment.

Chapter 3

A Bayesian hierarchical model for aggregating opinions by using predictions about the beliefs of others

This chapter mostly consists of material from a working paper with the same title that is joint work with Dražen Prelec.

Introduction

We begin this chapter by recalling that majority or confidence-weighted voting may lead to the wrong answer. To illustrate this, imagine that a committe is voting on two alternative launch strategies for a new product. Suppose that the default strategy appears superior on most dimensions, but has a subtle flaw that is evident to only a small fraction of committee members. These members will vote against it on grounds of prudence, even though they are not certain whether the flaw would prove fatal. By contrast, most members of the committee do not see the flaw in the default strategy and vote for it with high confidence. The majority therefore endorse the wrong strategy, and do so even if votes were weighted by confidence, because the true experts are relatively less confident about the merits of the two proposals.

Chapter 2 provided a theoretical diagnosis of this problem and proposed a solution.

Standard crowd wisdom methods can, in principle, measure the actual distribution of opinions to a high level of precision. However, to correctly interpret the meaning of the actual distribution, one needs to know what distributions to expect under different hypotheses. Standard methods therefore miss a crucial piece of information, namely counterfactual distributions of opinions. Our possible world model in Chapter 2, shows that information about counterfactual distributions may be indirectly obtained by eliciting respondents' predictions of the distribution of opinions in the sample. An ideal Bayesian respondent would compute such a predicted distribution as a weighted average of opinions expected under different hypotheses, with the weights equal to her posterior probabilities over these hypotheses. If these predictions are consistent with a Bayesian model of belief updating, one can prove for the special case of two mutually exclusive alternatives, that the best alternative is not the one with the most votes, nor the one endoersed with most confidence, but rather the alternative that receives more votes than predicted. We called this selection mechanism the 'surprisingly popular' algorithm.[1]

As shown in Chapter 2, the surprisingly popular algorithm can be extended to multiple choice questions, and is guaranteed to select the best answer under very general conditions, as the sample size goes to infinity. The empirical evidence that we discussed suggests that the surprisingly popular algorithm can robustly outperform unweighted or weighted averaging of opinions, even without additional statistical modeling . However, the surprisingly popular algorithm has several limitations.

First, it outputs only a single most likely answer, but one would prefer to obtain a complete posterior probability distribution over answers and, in the multiple choice case, rank answers according to posterior likelihood. Second, the surprisingly popular algorithm assumes that respondents are noiseless Bayesians, whereas the votes and vote predictions of actual respondents are likely to contain noise. Third, the surpris-

[1]We can illustrate the use of the suprisingly popular algorithm on the above example. In the above example, committee members voting for the default strategy, who are unaware of the flaw, are likely to predict that almost everyone will share their view, and vote for the default. Those who vote against the default expect to be in the minority, as the flaw is subtle and not likely to be recognized. Thus, the percent of votes for the default strategy will fall short of expectations, while the percent voting for the alternative will exceed expectations. The surprisingly popular algorithm will thus correctly reject the default alternative.

ingly popular algorithm does not allow for other kinds of information to be easily incorporated, for example when external information, such as base rate frequencies, is available. Fourth, the surprisingly popular answer does not apply across multiple questions, and so potentially discards information that could be used to identify individual respondent expertise.

To address these limitations, we develop a generative possible world model (GPWM) that models how the vots and vote predictions that people give are generated and so casts the aggregation problem as one of statisitcal inference. The GPWM yields a posterior probability distribution over world states, rather than outputting only the most likely answer. Unlike other Bayesian hierarchical models, the GPWM can be applied at the level of a single question without requiring data about a respondent's historical accuracy. When respondents answer multiple questions, the GPWM can be applied across questions and used to infer respondent expertise. The GPWM models noise in the votes and vote predictions of respondents, and allows for external information to be easily incorporated. Unlike market-based mechanisms, it is not restricted to events with outcomes that can be described by a precise contract and since the inputs used by the GPWM are the same as those required by the Bayesian Truth Serum (Prelec, 2004), honest answers can be easily incentivized. By developing the GPWM, we show how to combine statistical aggregation techniques with predictions about others, suggesting new future directions.

In the rest of this chapter, we motivate and formalize the GPWM. We describe a version of the model that can be applied to single questions and then extend it to learn the expertise of respondents answering multiple questions. For context, we compare our model to the Bayesian cultural consensus model (Karabatsos and Batchelder, 2003; Oravecz et al., 2014, 2013) and a Bayesian cognitive hierarchy model (Lee and Danileiko, 2014). The GPWM shows good performance on a range of domains, in inferring both correct answers and the expertise of respondents. We conclude by discussing assumptions of the GPWM and possibilities for how to extend it.

3.1 The generative possible world model

3.1.1 The generative possible world model for single questions

We present the generative possible world model (GPWM) for single questions, and in the next section discuss a version for multiple questions that includes a parameter for respondent expertise. We present a model for binary questions, and later consider alternatives and extensions, including to non-binary multiple choice questions.

3.1.1.1 Ideal Bayesian respondents with common knowledge of the possible world model (PWM), but asymmetric information

The following model (Prelec et al., 2017), which was presented in Chapter 2, underlies the surprisingly popular algorithm, and is described here for the case of binary questions, two signals, and a cutoff voting rule. Suppose that N respondents answer a binary question. Each answer to the question corresponds to a world state $\Omega \in \{A, B\}$, and the correct answer, given the available evidence, corresponds to the actual world state denoted Ω^*. Respondent r receives a private signal T^r, but does not know the actual world state. There are the same number of possible signals as there are possible world states with $T^r \in \{a, b\}$ for binary questions. Respondents are exchangeable except for the signal that they receive. Respondents have common knowledge of the joint distribution of worlds and signals $p(\Omega, T^r)$, referred to as the possible world model (PWM).[2] The PWM is unconstrained (except that to avoid degeneracies $p(\Omega = I) > 0$, $p(T^r = i) > 0$ for all worlds I and signals i). The probability of respondent r receiving signal k is given by $p(T^r = k|\Omega = \Omega^*) = p(\Omega = \Omega^*, T^r = k)/p(\Omega = \Omega^*)$. An ideal Bayesian respondent r receiving signal k has two kinds of beliefs. First, a posterior distribution over worlds, computed by Bayes rule, so that their belief that the world is in state j is

$$p(\Omega = j|T^r = k) = p(T^r = k|\Omega = j)p(\Omega = j)/p(T^r = k).$$

[2]In formally describing the models in this chapter, we use the term 'possible world model' (PWM) to refer only to this joint distribution, rather than any of our other assumptions about how respondents vote and predict the votes of others.

Second, a posterior distribution over the signals received by an arbitrary respondent s, computed by marginalizing over possible worlds, so that their belief that respondent s received signal j is

$$p(T^s = j|T^r = k) = \sum p(T^s = j|\Omega = i)p(\Omega = i|T^r = k).$$

Given these two posterior distributions, respondent r provides two kinds of information to the aggregator. First, a vote $V^r \in \{A, B\}$ for whichever world has maximum probability under their posterior distribution over worlds

$$V^r = \arg\max_{j \in \{A,B\}} p(\Omega = j|T^r = k)$$

Second, a prediction of the fraction of respondents who vote for each option, which is equivalent to the probability of an arbitrary respondent voting for each option. This can be obtained by summing over the probabilities of signals which would lead to a respondent voting for a particular option. The prediction of the fraction of respondents voting for option A is thus

$$p(V^s = A|T^r = k) = \sum_{j \in \{a,b\}} p(T^s = j|T^r = k)p(V^s = A|T^s = j)$$

where the probability that a respondent who receives signal j votes for option A is

$$p(V^s = A|T^s = j) = \mathbb{1}\{p(\Omega = A|T^s = j) > 0.5\}.$$

Unlike many standard aggregation models that put constraints on the PWM such that that the consensus answer is correct (Austen-Smith and Banks, 1996), the PWM in the model above is unconstrained. Hence, the correct answer cannot be otained from only vote frequencies, or from the full posterior distribution over worlds from every respondent. The solution presented in Chapter 2, is to additionally elicit from respondents their predictions about the votes of others, and select the answer which has an actual vote frequency exceeding the vote frequency which the group predicts.

We now build on this model to overcome the limitations described above of simply selecting the surprisingly popular answer.

3.1.1.2 The generative possible world model (GPWM)

The generative possible world model (GPWM) is a Bayesian hierarchical model (Rossi et al., 2005), specifically a probabilistic generative model (Jordan, 2004; Kollar and Friedman, 2009). The graphical model (Jordan, 2004; Kollar and Friedman, 2009) for the single question GPWM is shown in Figure 3-1. The graphical model shows how unobserved latent parameters (unshaded nodes) are sampled and how observed votes and vote predictions (shaded nodes) are generated for each respondent. The GPWM maintains the structure of the model above, with Bayesian respondents who have common knowledge of the PWM, receive a signal, and formualte posterior distributions over both worlds and the votes of others. The GPWM makes three changes to the model above: (1) abstract distributions are replaced with parametric distributions, (2) a sampling process for the PWM is specified, (3) a noise model for votes and vote predictions is introduced, along with the appropriate noise parameters. The complete GPWM for single questions is summarized in statistical notation at the end of this section.

3.1.1.3 Parametric distributions

To enable statistical inference, a parametric form for each distribution in the GPWM must be specified. To specify a parametric form for the PWM, it is decomposed into a prior over worlds and a set of distribution over signals conditional on the world state. For binary questions, the prior over worlds is given by $\Psi = (\Psi_0, 1 - \Psi_0)$ where $p(\Omega = A) = \Psi_0$. The actual world is sampled from a categorical distribution with this prior over world states. We refer to the distribution of signals conditioned on world states as the signal distribution S, and represent it in the binary case as a 2×2 left stochastic matrix where the iI-th entry gives the probability of an arbitrary respondent receiving signal i when $\Omega^* = I$ (fixing an ordering of world states and signals). In other words, each respondent's signal is sampled from a categorical

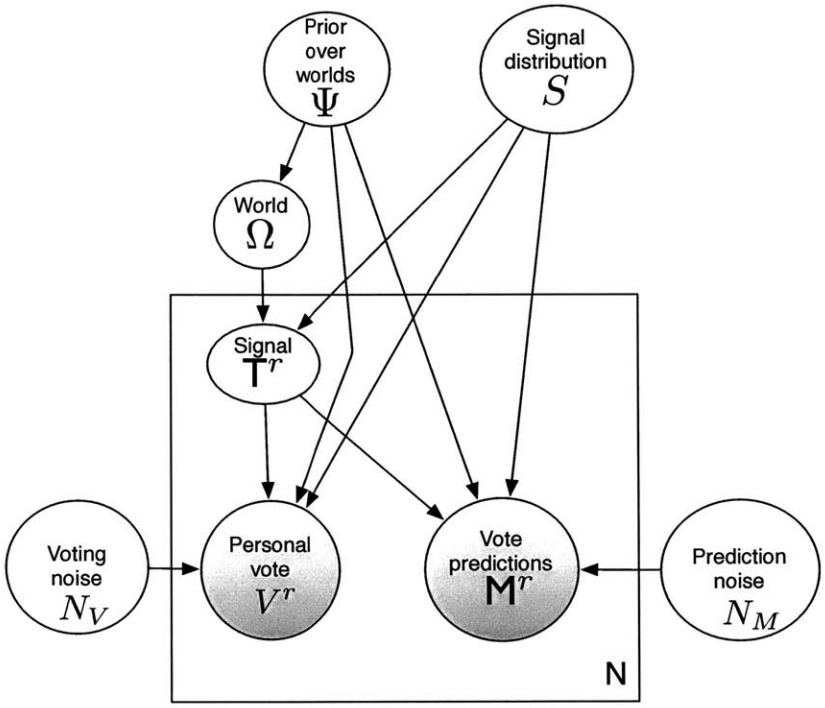


Figure 3-1: The single question generative possible world model (GPWM) shown using plate notation. This model of how votes and vote predictions are generated is used to infer a posterior distribution over latent variables, including the correct world state, given observed data from respondents. Nodes are random variables, shaded nodes are observed, an arrow from node X to node Y denotes that Y is conditionally dependent on X, a rectangle around variables indicates that the variables are repeated as many times as indicated in the lower right corner of the rectangle (Jordan, 2004; Kollar and Friedman, 2009).

distribution with the probabilities of the signals given by the Ω^*-th column of the signal distribution matrix. If $\Omega^* = I$, then we refer to signal i as the correct signal.

3.1.1.4 Sampling a PWM

To sample the PWM, we sample a world prior and a signal distribution independently. The prior probability that the world is in state A, i.e. Ψ_0, is sampled from a $Beta(1, 1)$ distribution. The signal distribution is sampled uniformly from the set of left stochastic matrices obeying the constraint that $S_{ii} > S_{ij}$ for all $i, j \neq i$. That is, we impose the constraint that $p(T^r = i|\Omega = I) > p(T^r = i|\Omega = J)$ for all signals i and worlds $I, J \neq I$. This generates a signal distribution such that signal i is more probable in world I than in any other world and imposes a constraint on the world in which a given signal is more likely, not a constraint on which signal is more likely for a given world. This does not imply that the majority of respondents receive a signal corresponding to the correct answer. That is, this prior over signal distributions guarantees that signal i is more likely to be received in world I than in world J, but allows signal i to be less likely than signal j in world I. For example, such a signal distribution may put probability 0.8 on signal a in world A and probability 0.7 on signal a in world B and thus have signal a be more likely than signal b in both worlds, leading to an incorrect majority if the actual world is B. The constraint that we place on the signal distribution allows one to index signals and world states and makes the model identifiable, analogous to imposing an identifiability constraint to alieviate the label switching problem when doing inference on mixture models (Diebolt and Robert, 1994; Jasra et al., 2005).

3.1.1.5 Noisy voting

A respondent's vote is modeled as a softmax decision function of their posterior distribution over world states, with the temperature of the softmax function given by a voting noise parameter N_V. The voting noise parameter is sampled from a $Gamma(3, 3)$ distribution. The parameters of this prior distribution were fixed in advance of running the model on the datasets.

3.1.1.6 Noisy vote predictions

Respondents give their prediction M^r of the fraction of the group who vote for each answer. For binary questions, M^r is completely specified by the fraction of people predicted to vote for option A. A respondent's prediction of this fraction is sampled from a Normal distribution truncated to the unit interval, with a mean given by their Bayesian posterior distribution on an arbitrary respondent voting for A and variance N_M, a prediction noise parameter that is identical for all respondents, and which is is sampled uniformly from $[0, .5]$.

3.1.1.7 Forward sampling and inference

We now summarize how votes and vote predictions are generated according to the GPWM. A PWM, that is a prior distribution over worlds and a signal distribution, is sampled as specified above, a world state is sampled conditional on the world prior, and the world and signal distribution are used to sample a signal for each respondent. Noise parameters are sampled from the specified priors. Respondents compute posterior distributions over worlds and the votes of others, and these are used to noisily sample votes and vote predictions for each respondent, given the noise parameters. For concreteness, we summarize below the complete GPWM using statistical notation. For notational convenience, rather than denoting worlds by $\{A, B\}$ and signals by $\{a, b\}$ we represent both worlds and signals by $\{0, 1\}$ so that we do not require categorical distributions. We introduce notation for several matrices that can be computed by any respondent, given their common knoweldge of the PWM. The matrix F gives the marginal distribution over signals, W gives the posterior probabilities over worlds conditional on each signal being received ($W_{i,j} = p(\Omega = i|T^r = j)$), and X gives the posterior distribution over signals conditional on the signal that was received ($X_{i,j} = p(T^s = i|T^r = j)$). The GPWM specifies the following generative

process:

$$
\begin{aligned}
\Psi_0 &\sim Beta(1,1) \\
\Psi &= (\Psi_0, 1-\Psi_0) \\
\Omega^*|\Psi &\sim Bernoulli(\Psi_1) \\
S' &\sim Beta(1,1) \\
S'' &\sim Beta(1,1) \\
S &= \begin{bmatrix} \max(\{S', S''\}) & \min(\{S', S''\}) \\ 1-\max(\{S', S''\}) & 1-\min(\{S', S''\}) \end{bmatrix} \\
T^r|S, \Omega^* &\sim Bernoulli(S_{1,\Omega^*}) \\
F &= S\Psi \\
W &= \begin{bmatrix} S_{0,0}\Psi_0/F_0 & S_{1,0}\Psi_0/F_1 \\ S_{0,1}\Psi_1/F_0 & S_{1,1}\Psi_1/F_1 \end{bmatrix} \\
N_V &\sim Gamma(3,3) \\
V^r|N_V, W, T^r &\sim Bernoulli(\exp(W_{1,T^r}/N_V)/(\exp(W_{0,T^r}/N_V) + \exp(W_{1,T^r}/N_V)) \\
X &= SW \\
N_M &\sim Uniform([0,.5]) \\
M^r|, N_M, T^r &\sim Normal_{[0,1]}(X_{0,T^r}\mathbb{1}\{W_{1,0} > 0.5\} + X_{1,T^r}\mathbb{1}\{W_{1,1} > 0.5\}, N_M)
\end{aligned}
$$

Inference via Markov Chain Monte Carlo is discussed later, and uses the observed vote and vote prediction of each respondent to infer the value of all latent variables, including a posterior distribution over answers.

3.1.2 The generative possible world model for multiple questions

Figure 3-2 displays a version of the GPWM for data from respondents who answer multiple questions. It incorporates respondent-level expertise, learnt across Q questions, into the GPWM. For each question, a world prior, world, signal distribution,

and noise parameters are sampled as for single questions, with no relationship between these parameters across questions.

Respondent expertise in the GPWM does not reflect the noise in each respondent's answers. Instead, an essential difference relative to other statistical models of aggregation, is that respondent expertise corresponds to how likely respondents are to receive the correct signal, rather than how likely respondents are to err in reporting answers given their signal. We refer to this as 'information expertise' since it reflects differences in the information or insight that respondents have, not differences in how much noise they introduce when interpreting this information or insight. That is, an expert is someone who is likely to receive information diagnostic of the correct answer, rather than someone who introduces little garbling into the information that they received. The introduction of information expertise means that in the multiple question model every respondent no longer has an identical probablity of receiving a particular signal given by the signal distribution.

The information expertise parameter for respondent r is denoted I^r, and has a uniform prior distribution on $[0, 1]$. If $\Omega^* = J$, then the probability of receiving signal j increases linearly with the information expertise and the probability of receiving a different signal is decreased by the same amount. That is, if $S_{jJ} = p$ then $p(T^r = j|\Omega^* = J) = p + I^r(1 - p)$ and likewise the probability of receiving other signals is decreased by $I^r(1 - p)$. For example, suppose $S_{aA} = 0.4$. Then, if $I^r = 0$ then $p(T^r = a|\Omega^* = A, I^r) = 0.4$, if $I^r = 0.5$ then $p(T^r = a|\Omega^* = A, I^r) = 0.7$, and if $I^r = 1$ then $p(T^r = a|\Omega^* = A, I^r) = 1$. The information expertise parameter does not determine the accuracy of a respondent's answer in absolute terms, but rather relative to the question difficulty. For example, for an easy question where the signal distribution gives a high chance of receiving the correct signal, even if someone has an expertise of 0 they will likely still receive the correct signal.

This model of information expertise does not allow for respondents who are less likely to receive the correct signal than the probability given by the signal distribution. Of all respondents, the answers of those with expertise 0 are the most uninformative. If the model included expertise values between 0 and -1, then someone with expertise

-1 would receive a signal perfectly anti-correlated with the actual world and so provide the same information content as someone with an information expertise of 1.[3]

Respondents are modeled as formulating their personal votes and vote predictions without taking information expertise into account. That is, each respondent assumes that all respondents, including themself, have an information expertise of 0. In the next section, we discuss an extension to the GPWM that allows for respondents to know their own information expertise.

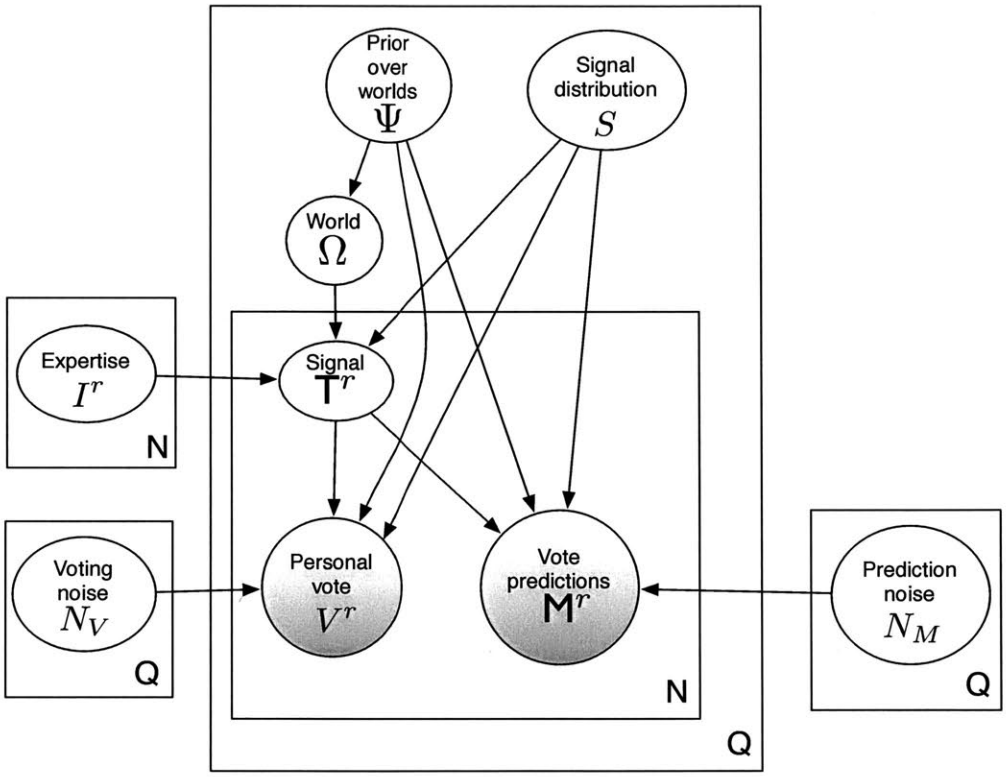


Figure 3-2: The multiple question generative possible world model (GPWM) which is applied across questions, with N respondents answering Q questions. It uses the single question GPWM, but includes information expertise which affects how likely an individual is to receive the correct signal.

[3]Restricting to positive expertise guards against inferring incorrectly that respondents consistently in the minority have expertise such that they receive incorrect signals, and allows the correct inference that these respondents are actually consistently correct, depending on their predictions.

3.1.3 Two extensions to the generative possible world model: confidence and expertise self-knowledge

3.1.3.1 Respondent confidence.

A respondent's confidence is modeled as a noisy report of their Bayesian posterior probability, conditional on their received signal, on their answer. That is, if respondent r received signal k and voted for answer j, their confidence is a noisy report of $p(\Omega = j|T^r = k)$. A respondent's confidence is assumed to be sampled from a Normal distribution (truncated between 0 and 1), with a mean given by their Bayesian posterior on their answer, and variance N_C a paramter that governs how noisy the confidence reports are of all respondents. For binary questions, respondents should give a confidence from 50% to 100%, since if someone had less than 50% confidence in an option they should have voted for the alternative option. We assume a truncated Normal distribution with support from 0, rather than 0.5, to allow for respondent error. The confidence noise N_C is sampled from a uniform distribution on $[0, .5]$. Confidence is incorporated both when applying the model to questions separately and when applying the model across questions.

3.1.3.2 Expertise self-knowledge.

A second extension to the GPWM assumes that respondents know their own information expertise, rather than simply assuming it is 0. If respondents assume that their information expertise is 0 then, although different respondents have different probabilities of receiving signal i because of their differing expertise, all respondents who do receive signal i have the same posterior beliefs over worlds and the signals of other respondents. Suppose instead that a respondent has accurate knowledge of their own information expertise. This implies that the posterior that they compute over worlds is

$$p(\Omega = i|T^r = j, I^r = e) = p(T^r = j|\Omega = i, I^r = e)p(\Omega = i)/p(T^r = j)$$

where

$$p(T^r = j|\Omega = i, I^r = e) = S_{ji} + e(1 - S_{ji}).$$

Assuming that respondents know their own expertise in this way has the effect that given two respondents receiving the same signal, the respondent who knows that she has high information expertise will put higher probability on the answer implied by the signal than will the respondent who believes that he has low information expertise. The respondents will also put different probabilities on the signals received by other respondents, since they have differing beliefs about the probability of different possible worlds. A significant disadvantage to modeling respondents as knowing their own information expertise is the increase in computation required for inference. Specifically, inference is done via Markov Chain Monte Carlo and if respondents know their own information expertise then the posterior distributions over worlds and signals, conditional on the received signal, have to be computed separately for every respondent at every inference step. Neither of these extensions improve performance on the seven studies we evaluate them on, as shown in Section 3.4.

3.2 Comparison models

For context and to gain insight into the factors affecting the performance of the GPWM in different settings, we compare the GPWM applied to questions separately to other methods that can be applied to individual questions: majority voting, selecting the answer that is more popular than the group predicts (Prelec et al., 2017), and the linear and logarithmic pools. We also compare the GPWM applied across multiple questions to other hierarchical Bayesian models that require multiple questions, specifically the Bayesian Cultural Consensus model (Oravecz et al., 2013, 2014) and a Bayesian cognitive hierarchy model (Lee and Danileiko, 2014).

3.2.1 Bayesian Cultural Consensus

Cultural Consensus Theory is a prominent set of techniques used to uncover the shared beliefs of a group (Romney et al., 1986; Batchelder and Romney, 1988; Weller, 2007). The theory deals with respondents answering a set of binary questions relating to the same topic. A respondent's answers are used to infer the extent to which each individual knows the culturally correct answers and the cultural consensus is determined by weighting more heavily the answers of culturally competent respondents. Unlike the GPWM, cultural consensus models cannot be applied to single questions and apply only to questions with binary answers. We specifically compare our model to a Bayesian hierarchical model called the Bayesian Cultural Consensus model (Oravecz et al., 2013, 2014).

The Bayesian Cultural Consensus model is applied to N respondents answering Q binary questions.[4] Respondents are indexed with r and questions with q. For each question q, a respondent r votes for either true or false, denoted by $Y_q^r \in \{0, 1\}$, and there is a culturally correct answer $Z_q \in \{0, 1\}$. For question q, a respondent r knows and reports Z_q with probability D_q^r and otherwise guesses true with probability g^r, a respondent specific guessing-bias. The competence D_q^r of respondent r at answering question q is given by the Rasch measurement model

$$D_q^r = \frac{\theta^r(1-\delta_q)}{\theta^r(1-\delta_q)+\delta_q(1-\theta^r)}$$

which is a function of the respondent's ability θ^r and the question difficulty δ_q. The competence of a respondent for a question increases with respondent ability, and decreases with the question difficulty. When respondent ability matches question difficulty, the probability of the respondent knowing the culturally correct answer for the question is 0.5. Uniform priors are assumed for all model parameters. The complete set of answers given by respondents is thus expressed as a probabilistic function of the culturally correct answer for all questions as well as the difficulty of each question, and the ability and guessing-bias of each respondent. The posterior distribution over

[4]We use consistent notation for all of the models in this chapter, departing from the notation originally used to describe a model if necessary.

the culturally correct answer for each question provides the aggregated answer of the group, and the inferred respondent ability and guessing-bias give information about each respondent's performance.

3.2.2 A Bayesian cognitive hierarchy model

Outside of cultural consensus theory, a number of hierarchical Bayesian models for aggregation have been developed that model how people produce their answers given latent knowledge, and then aggregate information at the level of this latent knowledge. For example, when playing "The Price is Right" game show, bids for a product may not correspond to knowledge of how much the product is worth (because of the competive nature of the show), and so aggregating at the level of latent knowledge rather than bids may give more accurate inferences (Lee et al., 2011b). Such hierarchical Bayesian models include models for aggregating over multidimension stimuli, for example combinatorial problems (Yi et al., 2012) and travelling salesman problems (Yi et al., 2010b). There are also a number of other statistical models, developed predominantly for crowdsourcing applications such as sentence annotation or image labeling, which have the same goal of aggregating information from multiple people (Raykar et al., 2010; Yan et al., 2011; Welinder et al., 2010; Whitehill et al., 2009; Bachrach et al., 2012; Kamar et al., 2015; Burnap et al., 2015). Modeling the cognitive processes behind someone's answer allows individual hetrogeneity to be estimated, for example the differing knowledge that respondents have in ranking tasks (Lee et al., 2012) or people's differing levels of noise and calibration when respondent's are answering binary questions using probabilties (Lee and Danileiko, 2014; Turner et al., 2014). We apply this Bayesian cognitive hierarchy model that incorporates noise and calibration Lee and Danileiko (2014) to the studies used in this paper, and compare it to the GPWM.

The Lee and Danileiko model (Lee and Danileiko, 2014) assumes N respondents answering a set of Q questions. Respondents are indexed by r, questions by q, and the answer of respondent r to question q is denoted Y_q^r. A respondent's answer gives their subjective probability that the world is in a particular state, and so $Y_q^r \in [0, 1]$. A

latent true probability π_q is associated with each question and two respondent-level parameters determine how respondents report this true probability. A respondent with calibration parameter δ_r perceives a probability $\psi_q^r = \delta_r \log(\frac{\pi_q}{1-\pi_q})$, which assumes a linear-in-log-odds calibration function. A respondent with expertise σ_r then reports a probability Y_q^r sampled from a Gaussian distribution centred around their perceived probabiltiy ψ_q^r, with variance given by the reciprocal of σ_r^2. That is, the larger σ^r, the more likely respondent r is to report a probablity close to their perceived probability. The paramters π_q and σ_r have a uniform prior distribution on the unit interval, and ψ_q^r has a $Beta(5,1)$ prior.

3.3 Evaluating the models

3.3.1 Data

We evaluate the models by reanalyzing the data from the seven studies discussed in Chapter 2. Three of these studies ("MIT class states", "Princeton states", and "MIT lab states") had respondents answer questions about state capitals. For the MIT lab states study, respondents gave their confidence of being correct on a scale from 50% to 100% and predicted the average confidence given by others. The prediction of the average confidence given by others is not used in the GPWM, but we return to this kind of prediction in the discussion section. Two studies ("Art professionals" and "Art laypeople") had respondents attempt to estimate the price of 20th Century artworks. In one study ("Lesions"), dermatologists judged whether lesions were benign or malignant, and in the last study ("Trivia") respondents evaluated the veracity of statements about trivia.

3.3.2 Applying the Generative Possible World Model

To apply the GPWM, the Metropolis-Hastings algorithm was used to perform Markov Chain Monte Carlo inference, with the signals marginalized out. Truncated Normal proposal distributions (centred on the current state with different fixed variances for

each parameter) were used for the world prior, noise, expertise and signal distributions (maintaining the constraint that the probability of signal a in state A is higher than in state B), and the opposite world state was proposed at each inference step. For inference across multiple questions, the question level parameters are conditionally independent across questions, conditioned on the respondent-level expertise values. We thus run MCMC chains for each question in parallel with the expertise parameter fixed, interspersed with an MCMC chain on only the expertise parameter. When applying the model to each question separately, 50000 Metropolis Hastings steps were used, 5000 of which were burn-in. When applying the model across multiple questions, 100 overall Metropolis-Hastings loops were used, the first 10 of which were burn-in, with each loop containing 2000 steps for the question parameters and 150 steps for the expertise parameter. For all three Bayesian hierarchical models, standard measures of autocorrelation and convergence were used to ensure that the samples from the model approximated the posterior.

3.3.3 Applying the Bayesian cultural consensus model

Cultural consensus models assume that there is a unidimensional answer key to the questions. A heuristic check that the data is sufficiently unidimensional is that the ratio of the first to the second eigenvalue of the agreement matrix is 3:1 or higher (Oravecz et al., 2014; Weller, 2007). We compute eigenvalue ratios for each study using respondents without missing data. Since for the lesions study, some lesions were seen only by one group we report the eigenvalue ratios separately for each group.This ratio is 2.76:1 for MIT class states, 2.62:1 for Princeton states, 3.32:1 for MIT lab states, 2.7:1 for Art laypeople, 8.92:1 for Art professionals, 2.81:1 for Trivia, 10.26:1 for the lesions with the 20 malignant and 40 benign split, and 6.73:1 for the lesions with the 40 malignant and 20 benign split. Most of the datasets are higher than the 3:1 standard for unidimensionality, and the model performed well on datasets not meeting this standard. The model learns a respondent-level guessing-bias towards true, and so for each study one of the answers is coded as true: the states studies and trivia study explicitly use true and false answers, in the art studies we coded true as

refering to the high price option (over $30 000), and in the lesions study we coded true as referring to the malignant option.

The Bayesian Cultural Consensus Toolbox (Oravecz et al., 2014) specicifies the model using the JAGS (Just Another Gibbs Sampler) model specification, which we edited to allow for unbalanced observational data, since not every respondent answered every question. In the MIT lab states study and the trivia study occassionally a respondent missed a question and in the lesions study only about half the respondents answered some questions due to the experimental design. To peform inference using the model, Gibbs sampling was run for 1000 steps of burn-in, followed by 10 000 iterations, using 6 independent chains and a step-size of two for thinning.

3.3.4 Applying the Bayesian cognitive hierarchy model

The cognitive hierarchy model requires subjective probabilities from respondents and so we apply the model only to data where this information is available, specifically the MIT lab states study, the lesions study, and the trivia study. The JAGS model specification provided by Lee and Danileiko with their paper was used, but edited to allow for unbalanced observational data. Gibbs sampling was run for 2000 steps of burn-in, followed by another 10000 iterations, using 8 independent MCMC chains.

3.4 Results

The GPWM is applied both to each question separately and across multiple question. The Bayesian cultural consensus model and cognitive hierarchy model are only applied across multiple questions. The cognitive hierarchy model is only applied to studies where confidences were elicited, whereas the other two models are applied to all seven studies. Results are shown with respect to the correct answer inferred for each question as well as the expertise parameters inferred for each respondent. The generative possible world model allows one to infer a prior over world states and this is compared to an empirical proxy for common knowledge about the likelihood of a city being a state capital.

3.4.1 Inferring the correct answers to questions

Of the three hierarchical Bayesian models discussed, only the GPWM can be applied to individual questions. There are other aggregation methods, however, that can be applied to individual questions. For all studies we compute the answer endorsed by the majority (counting ties as putting equal probability on each answer) and for studies where confidences were elicited we also compute a linear opinion pool given by the mean of respondent's personal probabilities (derived from votes and confidences), and a logarithmic opinion pool given by the normalized geometric mean of the probabilities that respondents assign to each answer (Cooke, 1991). We also show the result, from Chapter 2, of selecting the surprisingly popular answer.

Figure 3-3 shows the accuracy of the different methods at selecting the correct answer in terms of Cohen's kappa coefficient. Cohen's kappa is a standard measure of categorical correlation where a higher value indicates a higher degree of agreement with the actual answer. We compute it rather than the percentage of questions correct, since for some of the studies the relative frequencies of the different correct answers are unbalanced, which means that a method can have high percentage agreement even if it does not discriminate well and is instead biased towards the more frequent answer. Cohen's kappa is computed as $\kappa = \frac{p_o - p_e}{1 - p_e}$ where p_o is the relative observed agreement between the method and the actual answer, and p_e is the agreement expected due to chance, given the frequencies of the different answers output by the method.[5] A disadvantage of the kappa coefficient is that it does not take into account the probabilities output by a method, but only evaluates the answer that a method selects as most likely. Figure 3-4 shows the result of each method in terms of its Brier score. The Brier score evaluates probabilities, and a lower score indicates that a method tends to put high probability on the actual answer. There are a number

[5]There is not a standard approach to accomodate ties when computing the kappa coefficient for binary questions. In the case of ties, we construct a new set of answers with double the number of original answers. Answers which were not ties appear twice in the new set, and answers which were originally ties appear once as one answer and once as the other answer. The kappa coefficient is invariant to doubling the number of answers, and the standard error is a multiple of the number of original questions and so the doubling of the number of answers can be easily accounted for when computing standard errors.

of similar formulations of the Brier score which we compute as the average squared error between the probability given by a method and the actual answer. For methods that only output an answer, rather than a probability, we take the probabilities to be zero, one, or half in the case of ties.

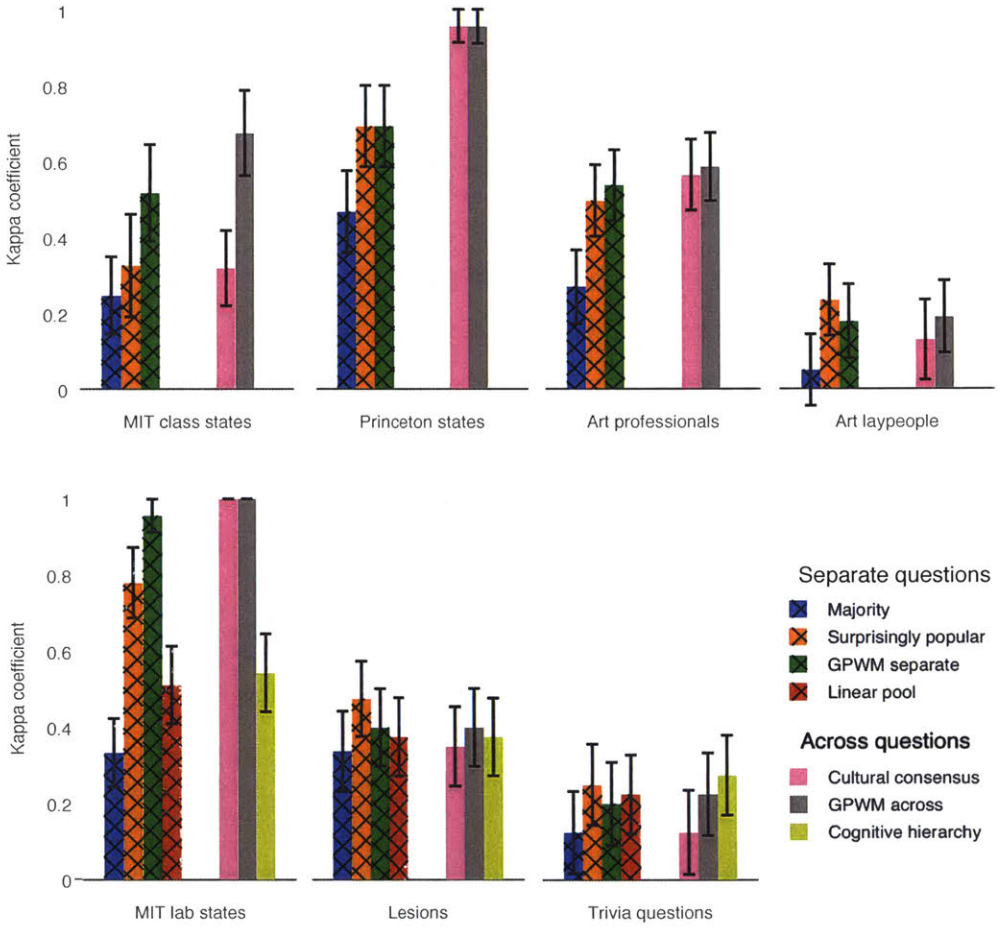


Figure 3-3: Performance of the various methods for aggregation on each dataset in terms of the kappa coefficient, with error bars indicating standard errors. The hatched bars show methods that are applied to single questions at a time. Confidences were only elicited in the studies shown in the bottom row.

We first consider methods that act on each question individually, shown with hatched bars in Figures 3-3 and 3-4. Studies where confidences were elicited are shown in the bottom row of each of these figures, and so studies in the bottom row show the result of more methods applied to them. The linear and logarithmic pools give

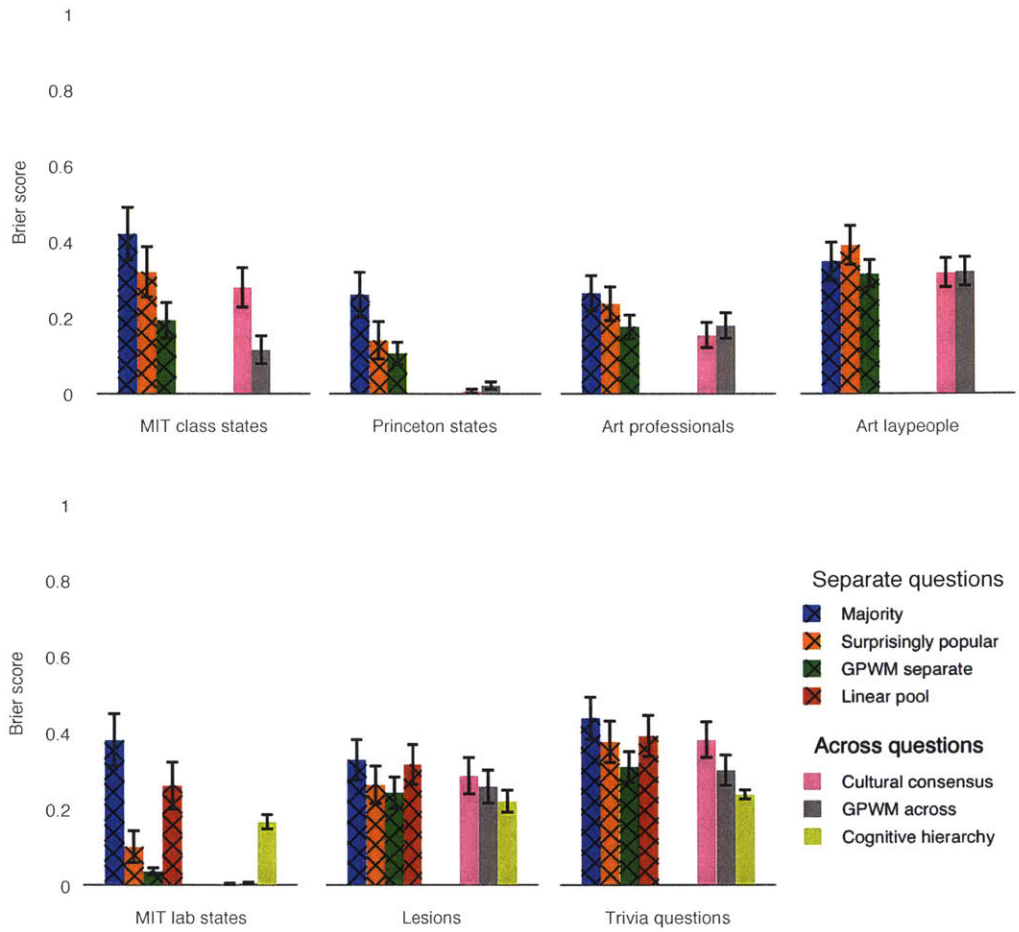


Figure 3-4: Performance of the various methods for aggregation on each dataset in terms of the Brier score, with error bars indicating bootstrapped standard errors. The hatched bars show methods that are applied to single questions at a time.

similar answers (the minimum kappa coefficient across studies when comparing the answers given by the linear and logarithmic pools is 0.86), and so we do not show the logarithmic pool results. Compared to the other methods that operate on questions individually, the GPWM applied separately to each question (which we refer to as GPWM separate) outperforms majority voting, and the linear and logarithmic pool across studies if we consider the accuracy of the answer selected by each method. This is displayed with respect to Cohen's kappa in Figure 3-3, but we can also compare the number of errors more directly. Across all 490 items, GPWM separate improved on the errors made by majority voting by 27% ($p < 0.001$, all p-values from a two-

sided matched pairs sign test on correctness unless otherwise indicated). Across the 210 questions on which confidences were elicited, GPWM separate improved on the errors made by majority vote by 30% ($p < 0.001$) and over the linear pool errors by 20% ($p < 0.02$). As can be seen, GPWM separate is able to uncover the correct answer even when applied to questions where the majority is incorrect. In terms of selecting one of two binary answers, the performance of the GPWM separate and the surprisingly popular answer is similar ($\kappa = 0.9$ across the 490 questions, and the answers are not significantly different by a two-sided matched pairs sign test, $p > 0.2$), which is to be expected since they build on the same model of respondent vote and vote predictions. However, GPWM separate has a better Brier score than the surprisingly popular answer for all studies, since it produces graded judgments rather than simply selecting a single answer. This ability to produce graded answers and reflect appropriate uncertainty in the inferences that can be drawn from respondent data illustrates one of the important advantages of developing statistical models over the surprisingly popular answer.

We also show the performance of models that require multiple questions (non-hatched bars in the figures): the Bayesian cultural consensus model, the cognitive hierarchy model, and GPWM applied across questions (which we refer to as GPWM across). GPWM across improves over GPWM separate ($p < 0.01$), although in terms of the kappa coefficient this improvement is small except for two of the states-capitals studies. The Bayesian cultural consensus model also shows good performance across datasets. It is similar to GPWM across in terms of the kappa coefficient, except for the MIT states-capitals study where its performance is not as good. This is also reflected in the difference between the correctness of the two methods in terms of absolute numbers of questions correct which only slightly favors GPWM across ($p = 0.057$). In a later section, we explore differences between GPWM across and the Bayesian cultural consensus model when there is not a consistent coding of answer options across questions. The cognitive hierarchy model selects similar answers to the linear pool ($\kappa = 0.92$) resulting in similar accuracy at selecting the correct answer, as measured by Cohen's kappa, but better performance than the linear pool with respect

to the probability it assigns to the correct answer, as measured by Brier score. The cognitive hierarchy model shows similar performance to the cultural consensus model and GPWM across on the trivia and lesions studies, but impaired performance on the MIT lab states study. This illustrates a general difference between GPWM across and the cognitive hierarchy model. In settings where confidence-weighted voting tends to be correct, the cognitive hierarchy model will perform well and may exceed the performance of the GPWM since it allows for differences in individual calibration. In settings where confidence-weighted voting tends to be incorrect, however, GPWM across has the potential to improve on the cognitive hierarchy model, with the caveat that this requires respondents to produce useful predictions about the answers of others.

We earlier discussed two possible extensions to the GPWM: incorporating confidences and assuming that respondents are aware of their own expertise. On the questions where confidences were elicited, we applied the GPWM with confidences incorporated both for questions separately and across questions. Applied to questions separately, the answers given by the GPWM with and without confidences were similar ($\kappa = 0.9$ on the selected answers, $r_s = 0.87$ on the returned probabilities). This was also the case when applying the GPWM across questions with and without confidences ($\kappa = 0.9$, $r_s = 0.86$). Hence, incorporating confidence made little difference to the GPWM results. We also applied GPWM across (both with and without confidences) assuming that individuals knew their own expertise. This again made little difference to the results for either the model without confidences ($\kappa = 0.9$ for answers, $r_s = 0.91$ for probabilities) or the model incorporating confidences ($\kappa = 0.9$ for answers, $r_s = 0.92$ for probabilities).

3.4.2 Inferring latent parameters: the world prior versus state capital mention frequencies

The GPWM allows one to infer the value of latent question-specific parameters other than the world state, such as the complete signal distribution and the prior over

worlds. The signal distribution gives information both about how people may vote in counterfactual world states, and about the actual distribution of information available to people. The world prior allows inferences about people's prior beliefs and what information is common knowledge amongst all respondents. The accuracy of these values is difficult to assess in general, but we analyze the inferred world prior in the state capitals studies. Previous work in cognitive science has demonstrated that in a variety of domains people have prior beliefs that are well calibrated with the actual statistics of the world (Griffiths and Tenenbaum, 2006). We use the number of Bing search results of the city-state pair asked about in each question (specifically the search query "City, State", for example "Birmingham, Alabama") as a proxy for the frequency of mentions of the city-state pair. For all three state capitals studies, the world prior on the named city being the capital (inferred from GPWM separate) has a moderate correlation with the Bing search count results under a log transform (MIT class: $r_S = 0.48, p < 0.001$, Princeton: $r_S = 0.49, p < 0.001$, MIT lab: $r_S = 0.55, p < 0.001$). This suggests that the GPWM inferences about the world prior may reflect common knowledge about how salient the named city is in relationship to the state.

3.4.3 Inferring respondent-level parameters

The GPWM applied across questions as well as the cultural consensus model and cognitive hierarchy model all infer respondent-level expertise parameters, and these inferences can be compared to the performance of individual respondents. Figure 3-5 shows the correlation of these respondent-level expertise parameters with the kappa coefficient of each respondent. For studies where confidences were elicited, the pattern of results is the same if respondent performance is measured using the Brier score.

Both the GPWM expertise parameter and the cultural consensus competence parameter show high correlation with individual respondent accuracy. For the 220 total respondents in all the studies, the correlation of respondent-level accuracy with GPWM expertise is r=0.79 and with cultural consensus competence is $r = 0.74$ (all respondent-level correlations are significant at the p<0.005 level). For the 97 respondents in the studies where confidences were elicited, the correlation of respondent-

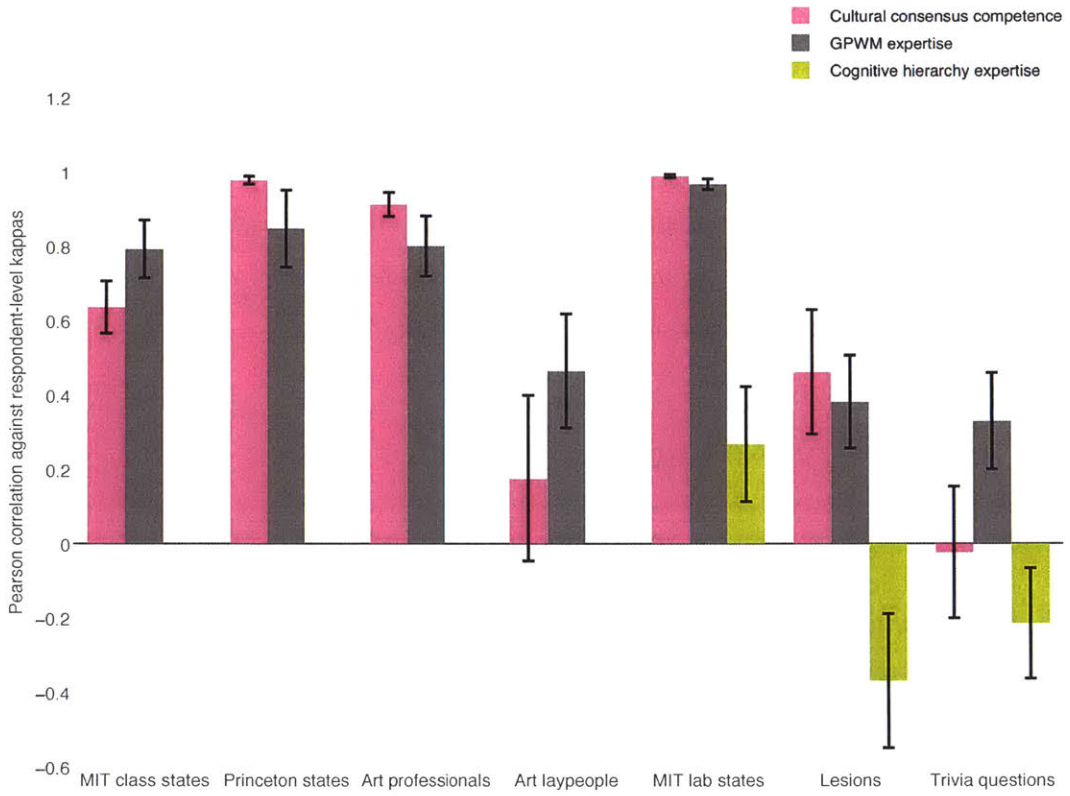


Figure 3-5: Pearson correlations of inferred respondent-level expertise parameters from each model against the accuracy of each respondent evaluated by their kappa coefficient. Error bars show bootstrapped standard errors.

level accuracy with GPWM expertise is r=0.78, with cultural consensus expertise is $r = 0.70$, and with cognitive hierarchy expertise is $r = 0.29$.

Two other pieces of information that may help predict a respondent's performance are the fraction of times that the respondent was in the majority, and the fraction of times that the respondent voted true. We examine the relationship between the expertise parameters and respondent performance if we partial out these two additonal factors. Across all studies, this partial correlation of respondent-level acurracy with GPWM expertise is $r = 0.85$ and with cultural consensus competence is $r = 0.74$. Across the studies where confidence was elicited, this partial correlation of respondent-level accuracy with GPWM expertise is $r = 0.76$, with cultural consensus competence

is $r = 0.70$, and with cognitive hierarchy expertise is $r = 0.34$.

For completeness, we also report on the other respondent-level parameters inferred by the models. The cultural consensus guessing-bias parameter has a high correlation with the fraction of questions for which a respondent says true ($r = 0.95$), but not with the accuracy of a respondent ($r = -0.11$, $p > .10$). The cognitive hierarchy model calibration parameter has a correlation of $r = 0.38$ with the kappa accuracy of respondents for the studies where confidence was elicited.

3.5 Factors affecting model performance

3.5.1 The consistency of answer coding across questions

Within the studies analyzed in this chapter, which answer was coded as true and which as false did not vary across questions. This need not always be the case, for example respondents could answer questions which sometimes ask whether a piece is worth more than $30,000 and sometimes ask whether a piece is worth less than $30,000. More generally, respondents may answer questions where there is simply no consistent ordering of the answer options across questions, for example questions asking which of two novel designs for a product will be more successful where the novel products in each question have no consistent labels.

To examine how the statistical models would fare in such situations, we constructed a half-reversed dataset where the questions in the first half of each study are coded in reverse, and another half-reversed dataset where the questions in the second half of each study are coded in reverse. To reverse a question, the correct answer to the question is swapped (i.e. true becomes false and vice versa) as is the answer of each respondent. Additionally, a respondent's prediction of the fraction of people voting true becomes their prediction of the fraction of people voting false. Figure 3-6 shows how the three Bayesian hierarchical models perform when applied on the half-reversed datasets in terms of average kappa coeffience.

The performance of GPWM across and the cognitive hierarchy model were not

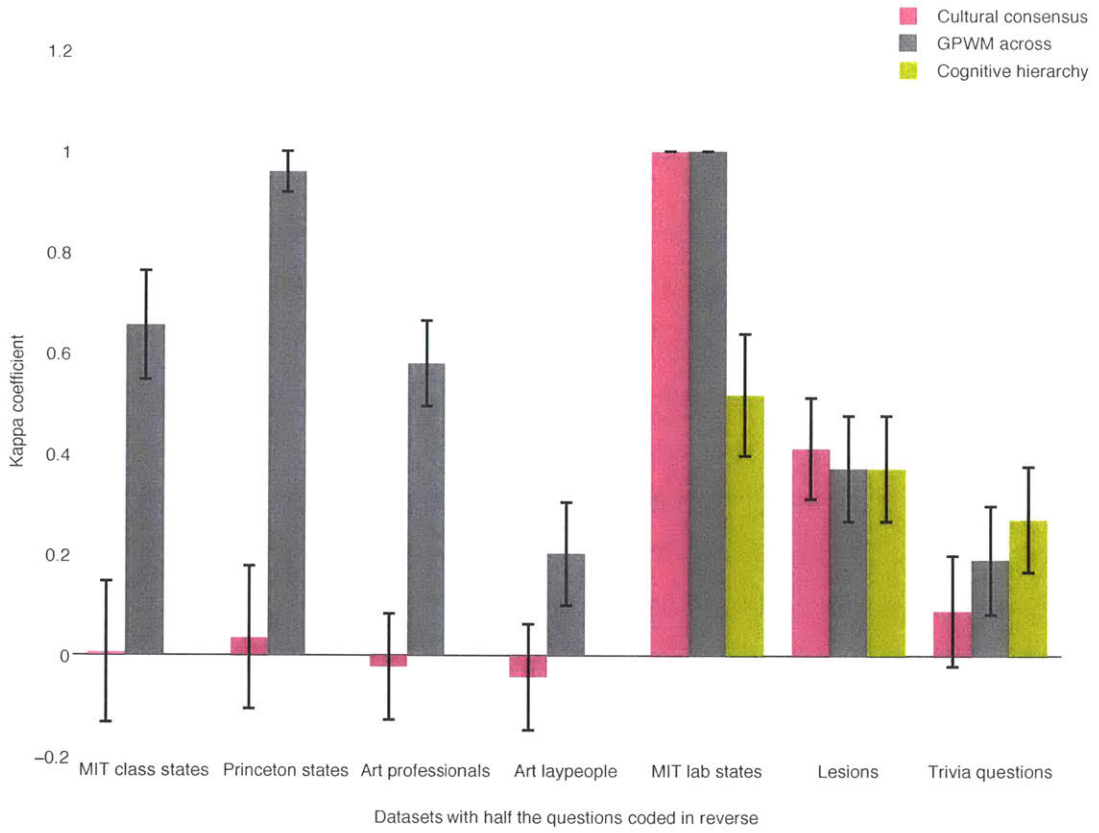


Figure 3-6: Performance of the three Bayesian hierarchical aggregation models when applied to the two half-reversed datasets. For each model, its kappa coefficient averaged across the two half-reversed datasets is shown. Error bars are standard errors.

affected by a non-consistent ordering of answer options across questions: the average of their kappa coefficients for the two half-reversed datasets were within 0.05 of the original kappas for every study. This was not the case for the Bayesian cultural consensus model, which had much lower performance on some of the studies when applied to the half-reversed datsets. In particular, for the MIT class states study, the Princeton states study, and both art studies it had a kappa of approximately 0 for the half-reversed datasets, in comparison to its good performance on the original datasets. This decrease in performance of the Bayesian cultural consensus model is because it relies, in part, on a respondent parameter which measures a bias towards guessing true. The other models, by contrast, do not represent such information and

so can perform well across questions that do not have a consistent coding of answer options.

3.5.2 The role of vote predictions

Since personal votes alone are insufficient to determine the correct answer if one does not assume that the correct signal is most probable, the GPWM also uses predictions about the votes of others. We evaluate the role of these predictions in the GPWM by lesioning the GPWM to ignore such predictions. When the GPWM is lesioned in this way, it infers answers that are very similar to those selected by majority voting. Across the seven datasets, the median spearman correlation between the posterior probability that the lesioned GPWM puts on an answer and the fraction of people voting for the answer is $r_S = 0.995$.

More generally, even in situations where such predictions are available, these can be more or less useful depending on how thoughtfully respondents make these predictions and how much variation there is across questions. If everyone always predicts that 50% of other resondents will answer true for every question these predictions will not be useful for improving the accuracy of the GPWM inferences.

3.6 Discussion

The GPWM depends on assumptions about: (1) the common knowledge that respondents share, (2) the signals that respondents receive, and (3) the computations that respondents make and how they communicate the results of these computations. We discuss each of these assumptions in turn, and the possible extensions to the GPWM that they suggest, and also discuss extending the GPWM to non-binary questions.

3.6.1 Knowledge shared by respondents

One set of modeling assumptions concerns the knowledge shared by respondents. Specifically, the GPWM assumes that respondents share common knowledge of the

world prior and signal distribution. However, neither of these assumptions are entirely correct: people do not exactly know these quantites, and beliefs about these quantities are not identical across people. As discussed when comparing the inferred world prior to Bing search results, knowledge of the world prior may reflect statistics of the environment that may be learnt by all respondents. In other cases, expert respondents may have a better sense of the world prior. For example, in diagnosing whether somebody has a particular disease based on their symptoms, knowledge of the base rate of the disease helps diagnosis but may not be known to everyone.

The GPWM could be extended to weaken the common prior assumption in various ways, although one could not simply assume that everyone had different belief about the prior over worlds since this would make the model non-identifiable. One could assume instead, for example, that respondents receiving the same signal share a common prior over worlds, but that respondents receiving different signals have different prior beliefs.[6] Alternatively, one could develop models where all respondents had noisy access to the actual world prior and signal distribution, and formulated beliefs about the knowledge that other respondents had about these quantities. For example, each respondent could sample from a distribution centred on the actual world prior and actual signal distribution. In the case of respondents answering many questions, the model could include parameters that governed the accuracy of a particular respondent's knowledge of these distributions. One could also incorporate domain knowledge available to the aggregator into the hyperprior over the world prior. For example if there is external knowledge of the base rate frequency of benign versus malignant lesions, a hyperprior could be chosen that would put more probability on a world prior that matched this base rate.

3.6.2 Signal structure

The GPWM assumes that there are the same number of signals as world states. It further assumes that the signals themselves have no structure: they are simply

[6]In this case, the model would require assumptions about what respondents believed about the prior possessed by respondents receiving other signals.

samples from $\{a, b\}$ (in the case of two worlds) with signal a more common in world A. This treats the information or insight available to a respondent coarsely in that it does not allow for more kinds of information available to respondents than there are answers to a question, or for respondents with different pieces of information to endorse the same answer. Models with more signals than worlds could be developed, requiring constraints on the worlds in which different signals were more probable. For example, signals could be generated from a hierarchical sampling process where each signal was a tuple with the first element indicating the world in which the signal was most likely and the second element indicating the rank of the probability of that signal amongst other signals with the same first element. That is, the signal distribution could be such that signal b_3, say, would be more likely in world b than in any other world, and of the other signals more likely in world b than in any other world it was the third most probable in world b. More generally, models with other kinds of assumptions about the signals that respondents receive could be developed to more faithfully model the information available to respondents when dealing with complex questions. For example, respondents could receive varying numbers of signals, a mix of public and private signals, or signals that are not simply categorical variables but rather have richer internal structure.

3.6.3 Respondent computations

The GPWM assumes that respondents compute Bayesian posteriors over answers and the votes given by others, and communicate the result of their computations without noise. An alternative is to model respondents as more plausible boundedly-rational agents, rather than simply as noisy Bayesians. For example, the Bayesian cognitive hierarchy model recognizes, based on work in the psychology of decision making (Zhang and Maloney, 2012), that respondents will be differentially calibrated with respect to the probabilities that they perceive and a similar calibration parameter could be included in the GPWM for the predictions of others answers. Modeling predictions of other people could also incorporate what is known about this process from social psychology. As just one example, as well as showing a false-consensus

effect, people sometimes also exhibit a false-uniqueness effect such that they do not take their own answer sufficiently into account when making their predictions (Chambers, 2008; Suls and Wan, 1987). Predictions of the answers of others could thus be modeled as resulting from a mixture of a respondent's prior over signals (i.e. their knowledge of the signal distribution) and their posterior over signals with the mixture weight given by a respondent-level false-uniqueness parameter.

In the version of the GPWM described in this chapter, we allow for noise in a respondent's vote, but do not assume that respondents take this noise into account when predicting the votes of others. Furthermore, respondents do not make predictions about the information expertise of other respondents when making their predictions about other people. We leave for future work the development of models where respondents take such factors into account when making predictions about others, for example by having respondents assume some distribution of information expertise across other respondents or by having respondents take into account the noise that they believe is present in the voting of other people.

Given the assumption that respondents compute a posterior distribution over world states and over the answers of others, additional statistics relating to either of these posteriors could be elicited and modeled. For example, in the MIT lab states study respondents were asked to predict the average confidence given by other respondents, which can be computed from these two posteriors. Such additional information could potentially help sharpen the model inferences. In practice, communicating such questions to respondents, respondent difficulty in reasoning about such questions, and respondent fatigue all impose constraints on the amount of additional data that could be usefully elicited in this manner. As well as eliciting other information from respondents, a different source of information is the actual answer to questions that have previously been asked and which have subsequently resolved. For example, if a group of respondents makes sales forecasts every quarter, the results of previous quarters will become known. By fixing the world state of such questions to their known actual value, the performance of the model on subsequent questions can be improved since the model will have better inferences about the information expertise of respondents.

3.6.4 Non-binary questions

Lastly, we discuss extending the GPWM to non-binary multiple choice questions. Most of this extension is straightforward, since the conditional distributions used in the model have natural non-binary counterparts. The world prior becomes a multinomial distribution, and the hyperprior a Dirichlet distribution, rather than a Beta distribution. The signal distribution for each world is likewise a multinomial distribution in the non-binary case. Given such distributions, respondents can compute a posterior distribution over worlds and the answers of others. Respondent votes can be modeled with a softmax decision rule given the Bayesian posterior over multiple possible worlds. Respondents can compute a Bayesian posterior distribution over the votes of others, but a model for adding noise to this posterior in the non-binary case is also required. One possibility is to sample from a truncated Normal distribution around each element of the posterior and then normalize the resultant draws to sum to 1, another is to sample from a Dirichlet distribution with a mean or mode determined from the Bayesian posterior and an appropriate noise parameter.

3.7 Conclusion

We have presented a generative possible world model that builds on the model of votes and vote predictions underlying selecting the surprisingly popular answer, while overcoming some of the limitations of the surprisingly populr answer. It shares some of the advantages of market-based mechanisms, for example it can be applied to single questions and can identify an expert minority, while sharing some of the advantages of existing statistical models, for example it can infer expertise across questions and does not suffer from the problems that can arise when individuals interact. The generative possible world model shows good performance both when applied to questions separately and when applied across multiple questions. It maintains this performance even when answers across questions do not have a consistent ordering. While the generative possible world model that we have proposed suggests multiple possible extensions and directions for future research, it is already a powerful method for

aggregating the beliefs of multiple individuals.

Chapter 4

Open-ended questions and richer predictions about others

This chapter discusses joint work with Dražen Prelec and Shane Frederick, and subsection 4.4.1 is joint work with Dražen Prelec and Sebastian Seung.

4.1 Introduction

Thus far in this thesis we have considered how asking people to predict the answers of others enables one to aggregate information from respondents answering categorical questions where the set of possible answers is known in advance. In this chapter, we extend our aggregation method to apply to questions where the space of answers is unknown prior to asking respondents for their opinions. In aggregating information about such questions, we will ask respondents for a richer kind of prediction about the opinions of others. We use these richer predictions to not only aggregate information about answer spaces that are unknown ex ante, but also hope to use them to shed more light on how people think about other people, especially people different from themselves. We use respondents' own answers and their predictions about the answers of others in a cognitive reflection test (Frederick, 2005) as a small case study to demonstrate these ideas.

4.2 Richer predictions

To aggregate information by selecting the surprisingly popular answer, we require predictions of the probability that other respondents will endorse each answer. As we have seen in Chapter 2, obtaining such probabilities is fairly straightforward when the answer space is simply a few possible options known in advance, as we can simply ask respondents to predict the percentage of respondents they believe will endorse each answer option. When this is not the case, we need to ask participants to make other kinds of predictions so that we can impute the necessary probabilities over the space of answers. When the answer space is categorical but unknown in advance, we propose asking participants to predict the three answers that they believe will be given most frequently by participants in the sample, and to predict what fraction of particpants will give each such answer that they predicted. We propose that to impute a respondent's prediction of the fraction of people endorsing other answers we simply partition each respondent's remaining prediction probability mass evenly amongst the other answers given by participants. For example, a respondent r may predict the three answers that they believe will be most commonly give, and further predict that 30% of respondents endorse the first of these, 25% endorse the second, and 20% endorse the third. If there are, say, two other answers that at least one respondent endorses, then r will be imputed to predict that 12.5% of respondents endorse the first of these, and 12.5% endorse the second.[1] To obtain more fine-grained information, one can increase the number of most frequent answers that people are asked to predict. Once the prediction probabilities are imputed, selecting the surprisingly popular answer can proceed as described in Chapter 2 for non-binary questions. As we will discuss, the predictions made by respondents who endorse different answers shed light on how respondents believe information or insight is distributed amongst the population.

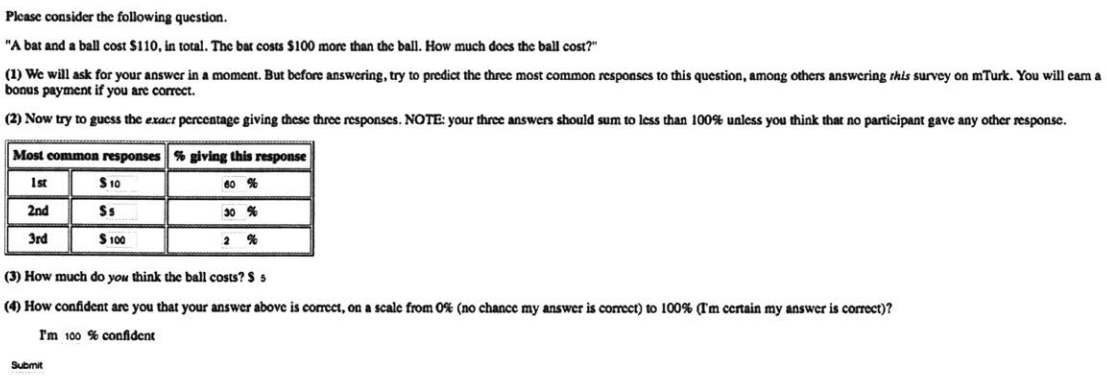
[1]Note that someone's imputed second-order predictions may not sum to 100% when they predict that people will endorse an answer that nobody actually predicts.

4.3 Cognitive reflection test case study

To illustrate the kinds of predictions that people give for open-ended questions, and how to use these predictions for aggregation we use (a long version of) the cognitive reflection test (CRT) (Frederick, 2005) as a case study. The test consists of questions which, for many people, cause an intuitive, incorrect answer to immediately spring to mind. People who do better on the CRT tend to be more patient, more risk seeking for losses, and more risk averse for gains (as reflected on hypothetical intertemporal choice tasks and in hypothetical choices between risky gambles). The CRT does not simply reflect IQ as measured by the Wonderlic Personnel Test, or self-reported SAT/ACT scores. The CRT is now widely used - to give just two examples, it is an excellent predictor of performance on a battery of tasks from the heuristics-and-biases literature beyond that offered by other measures of cognitive ability (Toplak et al., 2011) and people with higher CRT scores are more likely to play dominant strategies in the Beauty Contest game (Brañas-Garza et al., 2012).

4.3.1 Materials and methods

Participants ($N = 285$) were recruited from Amazon's Mechanical Turk service, restricted to people living in the United States. Participants were randomly assigned to one of ten questions from an extended cognitive reflection test (developed by Shane Frederick), with each question answered by at least 26 people. The questions in the extended cognitive reflection test have a similar flavor to that of the original test, with the first three questions the same as that of original cognitive reflection test Frederick (2005), except that the total price of the bat and ball question in the original test is \$1.10. One of the questions in the extended test ("Mary's mother had four children. The youngest three are named: Spring, Summer, and Autumn. What is the oldest child's name?") is particularly nicely illustrative of the open-ended nature of the domain since the answer is a word rather than a number. Participants could not preview the questions before opting into the experiment so as to prevent self-selection into answering particular questions.


Please consider the following question.

"A bat and a ball cost $110, in total. The bat costs $100 more than the ball. How much does the ball cost?"

(1) We will ask for your answer in a moment. But before answering, try to predict the three most common responses to this question, among others answering *this* survey on mTurk. You will earn a bonus payment if you are correct.

(2) Now try to guess the *exact* percentage giving these three responses. NOTE: your three answers should sum to less than 100% unless you think that no participant gave any other response.

Most common responses		% giving this response
1st	$ 10	60 %
2nd	$ 5	30 %
3rd	$ 100	2 %

(3) How much do *you* think the ball costs? $ 5

(4) How confident are you that your answer above is correct, on a scale from 0% (no chance my answer is correct) to 100% (I'm certain my answer is correct)?

I'm 100 % confident

Submit

Figure 4-1: An example of the elicitation procedure used in the cognitive reflection test study.

For the single question that they were considering, particpants were asked about their predictions about other people as well as about their own answer to the question. An example of the elicitation procedure used is shown in Figure 4-1 Participants were first asked to predict the first, second, and third most common answer to the posed question, given by other people answering the same question on Mechanical Turk. After giving their predictions of the most common answers, participants were asked to guess the exact percentage of the sample who gave each of these predicted answers. To make these instructions clear to participants, the table in which they had entered their answer predictions dynamically expanded to include a column in which they entered their predicted percentage for each answer. Participants were reminded that their predicted percentages should sum to less than 100%, unless they believed that no particpant gave any other response. Lastly, participants were asked for their own answer, and for their confidence in their answer on a scale from 0% ("no chance my answer is correct") to 100% ("I'm certain that my answer is correct"). Participants were paid a flat fee, and received a small bonus payment if their prediction of the most common answers was correct.

4.4 Results and analysis

After a technical diversion to show a result that we will use later in our analysis, we first discuss the kinds of predictions that respondents make, conditional on their own answers, and then discuss the use of respondent's predictions about others to uncover the correct answer to each question.

4.4.1 Technical diversion: The surprisingly popular answer for multiple-choice questions and the Bayesian Truth Serum

This subsection incorporates material from a working paper, Prelec, D. , Seung, S., and McCoy, J. (2015), "Identifying wisdom in the crowd by inferring implicit beliefs about possible worlds".

Later in this chapter, we use a particular formulation of the surprisingly popular answer for multiple-choice questions to aggregate answers, and additionally analyze the effect of weighting each respondents answers by the respondent's Bayesian Truth Serum score. In this subsection, we formalize the relationship between a formulation of the suprisingly popular answer and the Bayesian Truth Serum score. These technical details can be skipped if desired, we give a more intuitive motivation for weighting a respondent's answer by their Bayesian Truth Serum score in the next section.

Recall that in Chapter 2, Theorem 3 we consider the case of more than two worlds and the same number of signals as worlds, and show that if we assume that $p(a_i|s_i) > p(a_i|s_j)$ then the correct answer has the highest prediction-normalized vote. Given this assumption in Theorem 3, we now consider selecting the answer k which maximizes $p(a_{i^*}|v_k)$, that is $\arg\max_k p(a_{i^*}|v_k)$.

Theorem 4. *Assume $m = n$, $V(s_i) = v_i$, and $p(a_i|s_i) > p(a_i|s_j)$. Let a_{i^*} denote the actual world. Then, the vote for the correct answer is given by* $\arg\max_k p(a_{i^*}|v_k) = \arg\max_k \left\{\log p(v_k|a_{i^*}) + \sum_j w_j \log \frac{p(v_j|v_k)}{p(v_k|v_j)}\right\}$ *for any set of weights w_j satisfying $\sum_j w_j = 1$.*

Proof. From Bayes rule,

$$\frac{p(a_{i^*}|v_k)}{p(a_{i^*}|v_j)} = \frac{p(v_k|a_{i^*})}{p(v_j|a_{i^*})} \frac{p(v_j|v_k)}{p(v_k|v_j)}$$

After taking the logarithm and rearranging terms we have,

$$\log p(a_{i^*}|v_k) = \log p(v_k|a_{i^*}) + \log \frac{p(v_j|v_k)}{p(v_k|v_j)} + \log \frac{p(a_{i^*}|v_j)}{p(v_j|a_{i^*})}$$

We perform a weighted average over all j-s, and drop the rightmost term which does not depend on k, which yields,

$$\arg\max_k p(a_{i^*}|v_k) = \arg\max_k \left\{ \log p(v_k|a_{i^*}) + \sum_j w_j \log \frac{p(v_j|v_k)}{p(v_k|v_j)} \right\}$$

for any set of weights w_j satisfying $\sum_j w_j = 1$. □

Recall from Chapter 2, that the vote of respondent r is denoted V^r and that if respondent r received signal s_k then the prediction of respondent r about the probability of an arbitrary respondent q receiving signal s_j is denoted by $p(s_j^q|s_k^r)$. We let $x_k^r \in \{0,1\}$ indicate whether $V^r = k$ and $y_j^r = p(s_j^q|s_k^r)$. We define $\bar{x}_j$ as the fraction of respondents voting for j and$\bar{y}_j$as the geometric mean of the predictions of the fraction of respondents voting for j. If we let $w_j = \overline{x_j}$ and estimate $p(v_j|v_k)$ as the geometric mean of predictions of those who voted v_k we have

$$\arg\max_k p(a_{i^*}|v_k) = \arg\max_k \left\{ \frac{1}{n\bar{x}_k} \sum_r x_k^r u^r \right\}$$

where

$$u^r = \sum_s \sum_{k,j} x_k^r x_j^s \log \frac{\bar{x}_k y_j^r}{\bar{x}_j y_k^s} = \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_j \bar{x}_j \log \frac{y_j^r}{\bar{x}_j}$$

is the Bayesian Truth Serum (BTS) score (Prelec, 2004) of respondent r, which is a mechanism to incentivize respondents to answer honestly.

	Own answer correct	Own answer incorrect
Most frequent answer correct	46%	0%
Most frequent answer incorrect	28%	0%

Table 4.1: Fraction of times the correct answer was predicted to be the most common answer split up by questions where the most frequent answer was either correct or incorrect (rows) and by respondents depending on whether their own answer was either correct or incorrect (columns).

4.4.2 Predicting the answers of others

We begin by analyzing the answers that people predict that other people will give. Across all questions, respondents who correctly answered their CRT question listed the correct answer amongst the three that they predicted 82% of the time and listed the most frequent wrong answer 88% of the time. If we consider only the answer that they predicted as most frequent, this was the correct answer 40% of the time and the most common incorrect answer 42% of the time. Respondents who incorrectly answered their CRT question listed the correct answer amonst the three that they predicted 10% of the time and the most common incorrect answer 79% of the time. For half of the questions, nobody giving an incorrect answer predicted that the correct answer would be given as one of the three most frequent by other people, but this was as high as 44% of such respondents for one of the questions. As shown in Tables 4.1 and 4.2, we consider what answer people predicted as the most common and condition this on whether an incorrect answer was given more frequently than the correct answer for a given question. Respondents were more likely to predict the most frequent wrong answer as the most common when it actually did have the highest frequency, irrespective of whether the respondents themselves knew the correct answer. Respondents who did not know the correct answer never predicted that it would be the most common answer, although respondents who knew the correct answer often predicted it was the most common answer, although this happened less frequently for questions where the majority was incorrect.

	Own answer correct	Own answer incorrect
Most frequent answer correct	52%	37%
Most frequent answer incorrect	63%	44%

Table 4.2: Fraction of times the most frequent incorrect answer was predicted to be the most common answer split up by questions where the most frequent answer was either correct or incorrect (rows) and by respondents depending on whether their own answer was either correct or incorrect (columns).

4.4.3 Aggregating information in unknown answer spaces

We now compare aggregation methods for the open-ended CRT task that use predictions about others to those that rely solely on voting or confidence-weighted voting. We first compute for each question the surprisingly popular (SP) answer, which, as discussed throughout this thesis, essentially involves normalizing the frequency of votes for an answer by its predicted frequency of votes. Since there are more than two endorsed answers for each question, we use the non-binary version of the surprisingly popular answer. To compute the SP answer, we require a set of possible answers, vote frequencies for each answer, and vote predictions for each answer, conditional on the endorsed answer.

We take as the set of possible answers to a question simply the set of all answers that at least one participant gave as their own answer. Note that if an answer is predicted by some respondent as one of the most common answers, but no participant actually gives it as their own answer then it is not included in this set. We impute to each respondent a prediction over all of the possible answers by using the predictions they gave for the three answers they predicted to be most common, and distributing any remaining probability uniformly over the other answers, as discussed in Section 4.2. For 62% of respondents, there was no remaining probability mass since they predicted that everyone gave one of the three answers that they predicted to be most common. We discarded the 4% of respondents whose summed predictions for the three most common answers exceeded 100%.

For each question, we applied Theorem 4 (with $w_j = \overline{x_j}$ and geometric mean for predictions) to compute the suprisingly popular answer. We also computed which

answer had the majority vote, and which answer had the highest confidence-weighted vote. The surprisingly popular answer is correct for all ten questions, whereas majority voting and confidence-weighted voting both result in three errors. Whilst selecting the surprisingly popular answer is correct for more questions than the other two methods, this difference is obviously not significant due to the low power of a test with only ten questions. We thus turn to examining the clarity of verdict given by the different methods in more detail.

Computing the surprisingly popular answer as above is equivalant to selecting the answer that is endorsed by Bayesian respondents with the highest average Bayesian Truth Serum scores, as shown in Section 4.4.1. Instead of averaging the BTS scores of respondents who vote for the same answer, we can instead use differences between the predictions that respondents with the same vote give as a further source of information. That is, we regard the BTS score of a respondent as a measure of their expertise. We thus compute BTS scores for each respondent to each question. For each respondent, we have three measures associated with their vote: their BTS score, their confidence, and the frequency of their vote in the sample. We compare the strength of these measures at identifying respondents who gave the correct answer in two different ways.

First, we compare confidence-weighted voting against BTS-weighted voting. We computed an exponential weighting function, which resulted in a weight w_i for respondent i given as $w_i = e^{\beta s_i}$ where β is a free parameter and s_i is either their standardized confidence or standardized BTS score.[2] The normalized weight assigned to answer k is thus $(\Sigma_i w_i \mathbf{1}\{\mathrm{v_i} = \mathrm{k}\})/\mathbf{\Sigma_i} \mathrm{w_i}$ obtained by summing the weights of respondents whose vote v_i was for answer k, normalized by the sum of the weights of all respondents for that question. Figure 4-2 shows the result of a paired samples t-test between the normalized BTS weights on the correct answer versus the normalized confidence weights, across the ten questions. Irrespective of the choice of B shown in the figure, normalized weighted voting on the correct answer is higher for BTS than for con-

[2]So that BTS scores and confidences are on the same scale, giving neither an advantage, we standardize them by subtracting from the mean of respondents to the question, and dividing by the standard deviation. Whether or not we standardize makes little difference to the results.

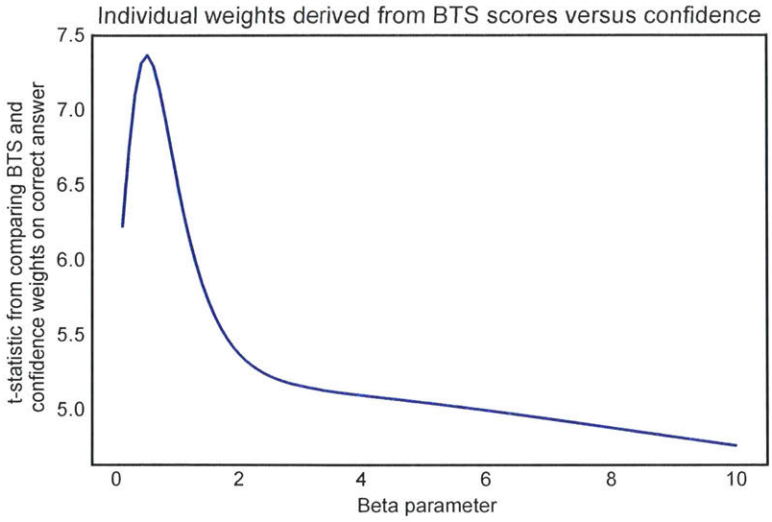


Figure 4-2: Results of weighting votes by BTS versus by confidence, using an exponential function that has a free parameter β as described in the text. A paired samples t-test was computed on the normalized weighted votes, using either BTS or confidence, assigned to the correct answer and averaged across questions. Note that for 9 degrees of freedom, a paired samples t-test is significant at the $\alpha = 0.05$ level when $t_9 = 1.833$.

fidence, indicating that the BTS score, which is obtained from a respondent's own answer and vote prediction, is a better measure of respondent expertise.[3]

Figure 4-3 shows how the weights derived from either BTS scores or confidence differ for individual questions for two representative values of β.[4] For these two representative values ofβ, the normalized BTS-weighted vote is higher than the normalized confidence-weighted vote on the correct answer for every question.

[3]Setting $\beta = 0$, would result in the same answers for confidence-weighted voting and BTS-weighted voting since this is equivalent to the fraction of respondents voting for each answer. At $\beta = 0$, the t-statistic is thus undefined since the standard deviation of the differences between the two methods is zero. At low weights, say $\beta = .1$, the weight BTS assigns to the correct answer is only slightly higher than the weight from confidence judgments (the average difference is 0.02 at $\beta = .1$) but since the standard deviation is so small this results in fairly large values of the t-statistic. The higher the choice of β, the more weight put on a few respondents with the highest scores.

[4]One could treat the normalized weights as probabilities and compute a Brier score for each method. This gives qualitatively similar results, with BTS weighting outperforming confidence weighting both averaged over questions and for individual questions.

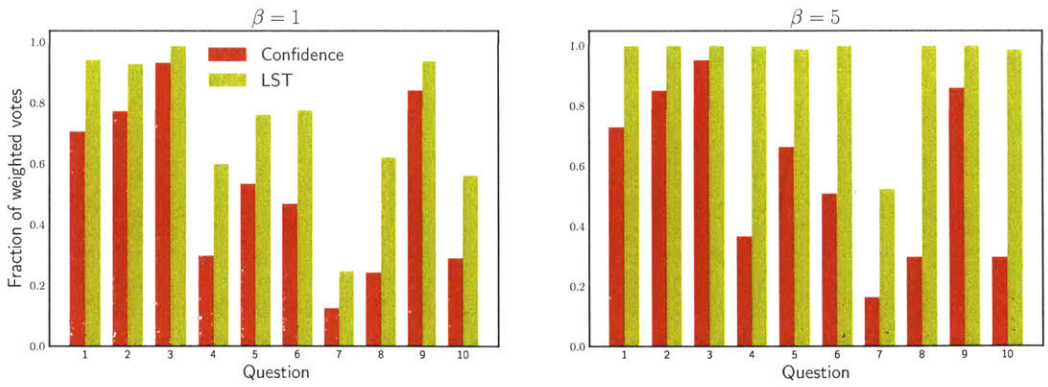


Figure 4-3: BTS-weighted voting versus confidence-weighted voting for individual questions given two representative settings of β in the exponential weighting function. BTS-weighted voting is higher for the correct answer than confidence-weighted voting for every question.

Second, we evaluate the mean accuracy of subsets of respondents selected according to three different expertise measures: the frequency of their answer in the sample, their confidence, and their BTS score. We selected subsets of respondents of every size 1 through 26 (the minimum number of respondents answering a question) adding respondents to a subset based on each of the three expertise measures being compared. That is, we constructed a subset of size 1 with the respondent having the highest BTS score, another such subset with the respondent having highest confidence, and a third subset of size 1 with the respondent giving the most frequent answer. Sometimes, especially for subsets based on the frequency of the answer, there were multiple respondents having the same score. In this case, we assigned everyone with the same score to the relevant subset. For example, if we were assigning subsets based on BTS score and four people were tied for the highest BTS score, then the subsets consisting of 1, 2, 3, and 4 people would all be the same and all contain these four people.

For each selected subset of respondents for each of the three measures, we computed the mean accuracy of respondents in the subset, averaging over all ten questions (Figure 4-4). The average accuracy of of subsets of respondents across the 26 sized subsets selected by BTS scores exceeds that of subsets selected by the confidence

of respondents or the typicality of their vote (paired sample t-tests give $t = 7.4$, $p < 0.001$ for BTS versus confidence subsets and $t = 7.6$, $p < 0.001$ for BTS versus typicality subsets).

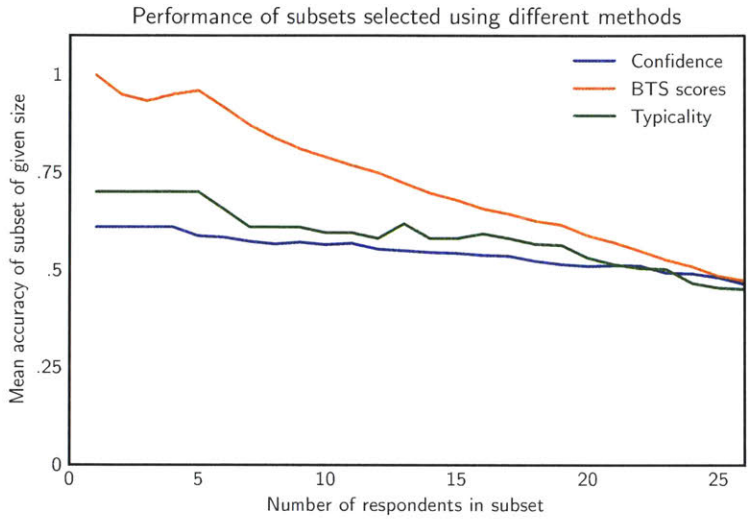


Figure 4-4: Mean accuracy of respondents in subsets selected on the basis of either frequency of answers, confidence, or BTS scores. A subset of each size is selected for each question, and results are shown averaged over all ten questions.

We can also show the same analysis for each question individually (4-5). Subsets of respondents selected on the basis of BTS scores consistently start with the highest accuracy and accuracy decreases as more respondents are included. Questions for which the majority is correct obviously start with accurate subsets when these subsets are selected on the basis of answer frequency. Confidence is not a reliable method of selecting respondents with expertise, as shown by the confidence column of Figure 4-5.

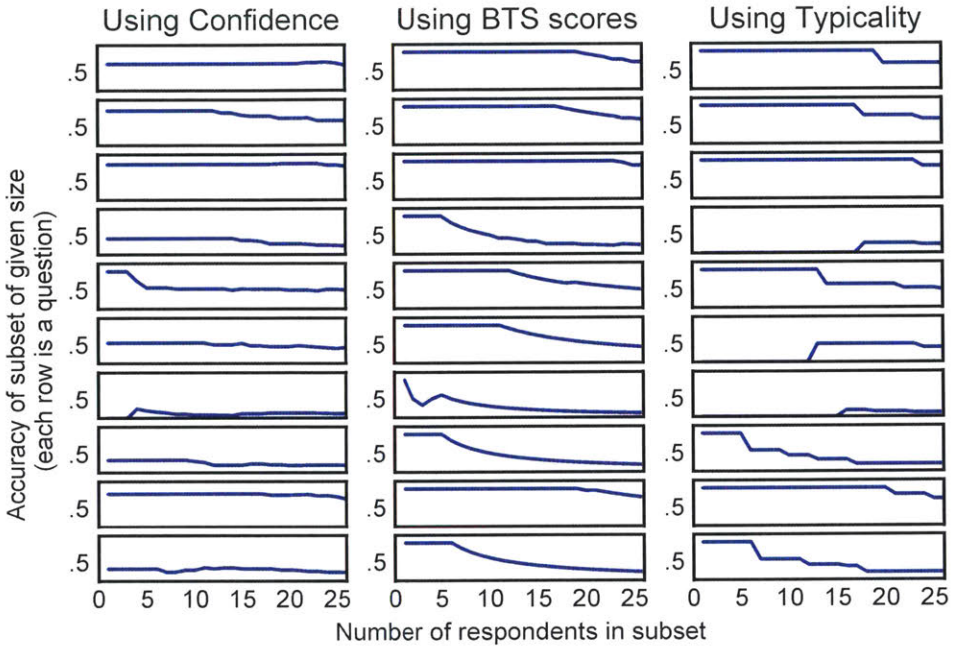


Figure 4-5: Mean accuracy of respondents in subsets selected using various methods, shown for each question individually. Each rown shows a single question.

4.5 Discussion

As can be seen from the results and analysis above, predictions about the answers of others give valuable information for aggregating answers to a cognitive reflection test. The cognitive reflection test questions are open-ended, and so we could not simply ask respondents to predict the distribution of people giving pre-specified answers, but rather asked them to predict the most common answers that people would give. The successful use of these predictions shows both that people were able to make such predictions and, moreover, that there was enough information in predictions about only the most common answers that the entire distribution was unnecessary. In section 4.4.2, which analyzed the predictions that people made as a function of the correctness of their own answer and whether the majority was correct, we saw that people who gave different answers to the questions (for example the correct answer

requiring deliberation versus the incorrect, intuitively salient answer) gave different predictions about the answers given by others. In particular, people who gave the correct answer were aware that many people would give a different answer to their own: the incorrect but intuitively salient answer.

There are a number of weaknesses of the study above, which could be addressed through future work. In particular, because we have only ten questions we are limited with respect to the amount of power we have. In this sample, the majority was correct for questions for which, in other samples, the majority is often incorrect Frederick (2005), making it more difficult to assess the advantage of vote predictions. Participants also answered only a single question, making it impossible to look for individual differences, or to identify how, for example, intelligence or the amount of deliberate thinking engaged in correlates with the ability to predict the answers of others. In future studies, particpants could answer multiplie questions making such an analysis possible.

The principle of selecting the surprisingly popular answer, examined throughout this thesis, depends on people making predictions that are informed by the signal that they receive. The question of what determines the predictions that different people make is particularly salient in this context, where the predictions are richer and where people differ in their reasoning abilities rather than differing in the information that they have. In keeping with Mata et al. (2013), we find that people who gave the correct answer are aware that others will fall for the intuitive, incorrect answer. Mata et al. (2013) assess this not through asking respondents to predict the answers that others will give, but rather through asking respondents to estimate solution rates and their own relative standing. In later work, respondents who gave the correct answer were found to be better at using meta cognitive cues (such as amount of thinking time) to predict the answers of others, since they were aware that deliberative thinkers gave a different answer to those giving the intuitive answer Mata and Almeida (2014). This was particularly the case for respondents who reported that they had considered an intuitive, incorrect answer when thinking about the problem. That respondents with the correct answer predicted a lower solution rate than respondents with an incorrect

answer was also reported in the original cognitive reflection test paper (Frederick, 2005). The idea is that, given a dual-process model of cognition, when people consider such problems an intuitive, incorrect answer comes quickly to mind since the problems are exactly chosen to lead to such decoy answers. Some people are able to override the temptation to give this intuitive answer and instead deliberate sufficiently to arrive at the correct answer. When making predictions about the answers of others, they are aware of the intuitive answer since they themselves experienced it, and so predict that some respondents will give it as an answer.

Under this account, the ability of respondents who gave the correct answer to give more accurate second-order predictions is not because they have superior social inference abilities in general but rather because they are aware of the intuitive but incorrect answer. More generally, when attempting to assess individual differences with respect to social inference skill, it is important to account for differences in predictions being due to differences in information.

An important question for future study asks about differences in predictions made by people who gave the same answer. Concretely, for example, what determines which respondents correctly recognize that they are in the minority when other respondents with the same answer predict that they are in a majority? One possibility is how much they considered alternative answers when thinking about the problem, another is differences in social inference ability. For the cognitive reflection test questions, many respondents can simulate or theorise about people who are not as deliberate as themselves, but we do not here have the evidence to infer how far downwards people can simulate or theorise. For example, given three levels of reasoning ability in a sample, people in the middle level may be better than people at the top level at predicting the answers of people in the bottom level. Future work could include, for example, having people answer multiple such cognitive reflection test type questions and asking them to predict the answers given by people of differing intelligence or with differing levels of performance on the cognitive reflection test. As another example, one could have chess players of differing levels of ability (as reflected in their elo ratings) make predictions about the moves of players of various elo ratings to

determine whether intermediate players are better than expert players at predicting the moves of beginners.

It is fitting that the cognitive reflection test case study appears at the end of this thesis since, although it is a nice illustration of the use of predictions to help aggregate information, it highlights how much there is still to learn about how people make predictions about others. Such future research will not only potentially help aggregate information but also grapple with the rich, complex, and important question of how people understand both themselves and others more generally.

Chapter 5

Concluding remarks

The primary contribution of this thesis has been to develop a method to aggregate judgments from multiple individuals, such that we can identify truth even when the majority is wrong. Standard methods such as voting, perhaps weighted by confidence, leave information on the table since they consider only the actual distribution of votes without accounting for vote distributions in other counterfactual worlds. We have developed a theoretical model of Bayesian respondents whereby predictions about the votes of others give information about such counterfactual distributions. We showed that, given this model, relying only on the posterior distribution over world states, as estimated by Bayesian respondents, is insufficient to reliably identify the correct world state. However, in the case of binary questions, selecting the surprisingly popular answer theoretically always identifies the correct answer. With the additional assumption that the respondents voting for an answer place the most probability on that answer, we show that normalizing votes by vote predictions also identifies the correct answer for non-binary questions. Across a number of domains, including trivia, geography, art, medicine, we show that selecting the surprisingly popular answer is more accurate, in practice, than standard methods, although we also discuss circumstances in which the method may fail, for example because of respondents exhibiting curse of knowledge effects.

Despite the advantages of selecting the surprisingly popular answer, this simple approach has a number of limitations. We addressed these limitations, using our

model of Bayesian respondents formulating their own answer and vote predictions, by developing a probabilistic generative model for information aggregation. This model provides a posterior distribution across world states and can be applied across multiple questions to identify respondents with high expertise. Using empirical data, we showed that this model has a number of advantages over other Bayesian hierarchical models for aggregation that do not include predictions of others. Both selecting the surprisingly popular answer and applying the probabilistic generative model require respondents to make predictions over a set of answers specified in advance. We extend our aggregation method to settings where this is not the case by using an elicitation procedure where people instead predict the most common answers that others will give, and succesfully apply this method to a cognitive reflection test as a case study.

The ability of people to give sensible predictions about the answers of others and the successful applications of selecting the surprisingly popular answer detailed in this thesis are encouraging, but there is much scope for future work building on this success. We take the opportunity here to briefly mention two areas of future research: practical applications of selecting the surprisingly popular answer, and investigating, informed by our model, how people make predictions about the answers of others together with using such predictions in contexts other than aggregation.

While selecting the suprisingly popular answer has been successful in the laboratory, how does it perform in more real-world situations? Future work (some of it already ongoing) will test the advantages of selecting the surprisingly popular answer for forecasting tasks with real-world events and experts. For the surprisingly popular answer to succeed in such contexts will require developing robust, efficient estimation methods for non-binary questions and methods to apply the surprisingly popular answer to continuous quantities. It will also be necessary to grapple with issues around applying the surprisingly popular answer when respondents have different incentives, respondents are in active communication, and institutions wish to know how much to trust the results. Most current work, including that in this thesis, deals with how to do aggregation at the level of people's answers to some question, but it would be desirable to aggregate the information or insights people have about some problem,

rather than simply their answers.

The theoretical model and empirical results discussed in this thesis raise many questions about how people make predictions, and how such predictions may be used. We give just three examples here to give a flavour. First, to what extent are people able to predict the answers of those with a different level of reasoning ability, for example are intermediate performers on a reasoning test better able than expert perfomers to predict the mistakes made by low performers? Second, what accounts for the false-uniqueness effects observed in our data, and how can this be moderated? Third, if people fail in systematic ways to predict the beliefs of people with answers different to their own on some question, can we use this to probabilistically detect when people are being deceptive about their own answer?

We began this thesis by pointing out that it is often important to be able to combine information from multiple people, especially when experts disagree. As the world becomes more complex and information increasingly distributed, this will only become more important. We hope that the ideas and methods contained in this thesis will help address this challenge.

Bibliography

Royce Anders and William H Batchelder. Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, 56(6):452–469, 2012.

Lisa R Anderson and Charles A Holt. Information cascades in the laboratory. *The American economic review*, pages 847–862, 1997.

Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. The promise of prediction markets. *Science*, 320(5878):877, 2008.

Alison Hubbard Ashton and Robert H Ashton. Aggregating subjective forecasts: Some empirical results. *Management Science*, 31(12):1499–1508, 1985.

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

David Austen-Smith and Jeffrey S Banks. Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review*, pages 34–45, 1996.

Yoram Bachrach, Tom Minka, John Guiver, and Thore Graepel. How to grade a test without knowing the answers - a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

William H Batchelder and A. Kimball Romney. Test theory without an answer key. *Psychometrika*, 53(1):71–92, 1988.

Joyce E Berg, Robert Forsythe, Forrest Nelson, and Thomas A Rietz. Results from a dozen years of election futures markets research. *Handbook of experimental economics results*, 1:742–751, 2008.

Joyce E Berg, George R Neumann, and Thomas A Rietz. Searching for google's value: Using prediction markets to forecast market capitalization prior to an initial public offering. *Management Science*, 55(3):348–361, 2009.

Eric Bonabeau. Decisions 2.0: The power of collective intelligence. *MIT Sloan management review*, 50(2):45, 2009.

Daren C Brabham, Kurt M Ribisl, Thomas R Kirchner, and Jay M Bernhardt. Crowdsourcing applications for public health. *American journal of preventive medicine*, 46(2):179–187, 2014.

Pablo Brañas-Garza, Teresa Garcia-Muñoz, and Roberto Hernán González. Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization*, 83(2):254–260, 2012.

David V Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 2014.

Alex Burnap, Yi Ren, Richard Gerth, Giannis Papazoglou, Richard Gonzalez, and Panos Y Papalambros. When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design*, 137(3):031101, 2015.

Colin Camerer, George Loewenstein, and Martin Weber. The curse of knowledge in economic settings: An experimental analysis. *The Journal of Political Economy*, pages 1232–1254, 1989.

John R Chambers. Explaining false uniqueness: Why we are both better and worse than others. *Social and Personality Psychology Compass*, 2(2):878–894, 2008.

Kay-Yut Chen, Leslie R Fine, and Bernardo A. Huberman. Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50(7): 983–994, 2004.

Yiling Chen, Chao-Hsien Chu, Tracy Mullen, and David M Pennock. Information markets vs. opinion pools: An empirical comparison. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 58–67. ACM, 2005.

Robert T Clemen and Robert L Winkler. Unanimity and compromise among probability forecasters. *Management Science*, 36(7):767–779, 1990.

Robert T Clemen and Robert L Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

Robert T Clemen and Robert L Winkler. Aggregating probability distributions. *Advances in Decision Analysis*, pages 154–176, 2007.

Marquis de Condorcet. *Essay sur l'application de l'analyse de la probabilité des decisions: redues et pluralité des voix.* l'Imprimerie Royale, 1785.

Roger M Cooke. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press, USA, 1991.

Roger M Cooke and Louis LHJ Goossens. Tu delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5):657–674, 2008.

David Court, Benjamin Gillen, Jordie McKenzie, and Charles R Plott. Two information aggregation mechanisms for predicting the opening weekend box office revenues of films: Boxoffice prophecy and guess of guesses. *Economic Theory*, 65(1):25–54, 2018.

Bo Cowgill and Eric Zitzewitz. Corporate prediction markets: Evidence from google, ford, and firm x. *The Review of Economic Studies*, 82(4):1309–1341, 2015.

Karen Croxson. Information markets for decision-making. *Prediction Markets: Theory and Applications*, 66:52, 2011.

Ely Dahan, Arina Soukhoroukova, and Martin Spann. New product development 2.0: Preference markets - how scalable securities markets identify winning product con-

cepts and attributes. *Journal of Product Innovation Management*, 27(7):937–954, 2010.

Ely Dahan, Adlar J Kim, Andrew W Lo, Tomaso Poggio, and Nicholas Chan. Securities trading of concepts (stoc). *Journal of Marketing Research*, 48(3):497–517, 2011.

Robyn M Dawes. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25(1):1–17, 1989.

Robyn M Dawes. False consensus effect. *Insights in decision making: A tribute to Hillel J. Einhorn*, page 179, 1990.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Jerker Denrell and Christina Fang. Predicting the next big thing: Success as a signal of poor judgment. *Management Science*, 56(10):1653–1667, 2010.

Jean Diebolt and Christian P Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375, 1994.

Theodoros Evgeniou, Lily Fang, Robin M Hogarth, and Natalia Karelaia. Competitive dynamics in forecasting: The interaction of skill and uncertainty. *Journal of Behavioral Decision Making*, 26(4):375–384, 2013.

Shane Frederick. Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4):25–42, 2005.

Simon French and David Rios Insua. *Statistical decision theory.* Wiley, 2000.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Francis Galton. The ballot-box. *Nature*, 75(1952):509–510, 1907a.

Francis Galton. One vote, one value. *Nature*, 75(1948):414, 1907b.

Francis Galton. Vox populi. *Nature*, 75:450–451, 1907c.

Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, pages 114–135, 1986.

Christian Genest, Samaradasa Weerahandi, and James V Zidek. Aggregating opinions through logarithmic pooling. *Theory and decision*, 17(1):61–70, 1984.

Daniel G Goldstein and Gerd Gigerenzer. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75, 2002.

Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773, 2006.

Bernard Grofman, Guillermo Owen, and Scott L Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15(3):261–278, 1983.

Robin Hanson. Shall we vote on values, but bet on beliefs? *Journal of Political Philosophy*, 2013.

Reid Hastie and Tatsuya Kameda. The robust beauty of majority rules in group decisions. *Psychological review*, 112(2):494, 2005.

Ralph Hertwig. Tapping into the wisdom of the crowd-with confidence. *Science*, 336 (6079):303–304, 2012.

Pamela J Hinds. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied*, 5(2):205, 1999.

Teck-Hua Ho and Kay-Yut Chen. New product blockbusters: The magic and science of prediction markets. *California Management Review*, 50(1):144–158, 2007.

Stephen C Hora, Benjamin R Fransen, Natasha Hawkins, and Irving Susel. Median aggregation of distribution functions. *Decision Analysis*, 10(4):279–291, 2013.

Daniel J Isenberg. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, 50(6):1141, 1986.

Ajay Jasra, Chris Holmes, and David Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

Leslie K John, George Loewenstein, and Dražen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

Michael I Jordan. Graphical models. *Statistical Science*, pages 140–155, 2004.

Victor Richmond R Jose, Yael Grushka-Cockayne, and Kenneth C Lichtendahl Jr. Trimmed opinion pools and the crowd's calibration problem. *Management Science*, 60(2):463–475, 2013.

Ece Kamar, Ashish Kapoor, and Eric Horvitz. Identifying and accounting for task-dependent bias in crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

George Karabatsos and William H Batchelder. Markov chain estimation for test theory without an answer key. *Psychometrika*, 68(3):373–389, 2003.

Jane Kennedy. Debiasing the curse of knowledge in audit judgment. *Accounting Review*, pages 249–273, 1995.

Norbert L Kerr and R Scott Tindale. Group performance and decision making. *Annu. Rev. Psychol.*, 55:623–655, 2004.

Marc Keuschnigg and Christian Ganser. Crowd wisdom relies on agentsś ability in small groups with a voting aggregation rule. *Management Science*, 2016.

Boaz Keysar, Linda E Ginzel, and Max H Bazerman. States of affairs and states of mind: The effect of knowledge of beliefs. *Organizational Behavior and Human Decision Processes*, 64(3):283–293, 1995.

Daphne Kollar and Nir Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

Asher Koriat. When are two heads better than one and why? *Science*, 336(6079): 360–362, 2012.

Joachim Krueger and Russell W Clement. The truly false consensus effect: an ineradicable and egocentric bias in social perception. *Journal of personality and social psychology*, 67(4):596, 1994.

Krishna K Ladha. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, pages 617–634, 1992.

Richard P Larrick, Albert E Mannes, Jack B Soll, and Joachim I Krueger. The social psychology of the wisdom of crowds. *Social psychology and decision making*, pages 227–42, 2011.

Maël Lebreton, Raphaëlle Abitbol, Jean Daunizeau, and Mathias Pessiglione. Automatic integration of confidence in the brain valuation signal. *Nature neuroscience*, 2015.

Michael Lee, Mark Steyvers, Mindy DeYoung, and Brent Miller. A model-based approach to measuring expertise in ranking tasks. In *Proceedings of the Cognitive Science Society*, volume 33, 2011a.

Michael D Lee and Irina Danileiko. Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9(3):259–273, 2014.

Michael D Lee, Shunan Zhang, and Jenny Shi. The wisdom of the crowd playing the price is right. *Memory & cognition*, 39(5):914–923, 2011b.

Michael D Lee, Mark Steyvers, Mindy De Young, and Brent Miller. Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science*, 4(1): 151–163, 2012.

Kenneth C Lichtendahl Jr, Yael Grushka-Cockayne, and Robert L Winkler. Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611, 2013.

Joseph Lipscomb, Giovanni Parmigiani, and Vic Hasselblad. Combining expert judgment by hierarchical modeling: An application to physician staffing. *Management Science*, 44(2):149–161, 1998.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

Spyros Makridakis and Robert L Winkler. Averages of forecasts: Some empirical results. *Management Science*, 29(9):987–996, 1983.

André Mata and Tiago Almeida. Using metacognitive cues to infer others' thinking. *Judgment and Decision making*, 9(4):349, 2014.

André Mata, Mário B Ferreira, and Steven J Sherman. The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of personality and social psychology*, 105(3):353, 2013.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5):1106–1115, 2014.

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

Peter A Morris. Combining expert judgments: A bayesian approach. *Management Science*, 23(7):679–693, 1977.

David G Myers and Helmut Lamm. The group polarization phenomenon. *Psychological bulletin*, 83(4):602, 1976.

Raymond S Nickerson. How we know - and sometimes misjudge - what others know: Imputing one's own knowledge to others. *Psychological bulletin*, 125(6):737, 1999.

Raymond S Nickerson. The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Science*, 10(5):168–172, 2001.

Tudor I Oprea, Cristian G Bologa, Scott Boyer, Ramona F Curpan, Robert C Glen, Andrew L Hopkins, Christopher A Lipinski, Garland R Marshall, Yvonne C Martin, Liliana Ostopovici-Halip, et al. A crowdsourcing evaluation of the nih chemical probes. *Nature chemical biology*, 5(7):441–447, 2009.

Zita Oravecz, Royce Anders, and William H Batchelder. Hierarchical bayesian modeling for test theory without an answer key. *Psychometrika*, pages 1–24, 2013.

Zita Oravecz, Joachim Vandekerckhove, and William H Batchelder. Bayesian cultural consensus theory. *Field Methods*, page 1525822X13520280, 2014.

Fumika Ouchi. *A literature review on the use of expert opinion in probabilistic risk analysis.* World Bank Washington, DC, 2004.

Charles R Plott and Kay-Yut Chen. Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. 2002.

Charles R Plott, Benjamin J Gillen, and Matthew Shum. A parimutuel-like mechanism from information aggregation: A field test inside intel. 2014.

Dražen Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695): 462–466, 2004.

Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

A Kimball Romney, Susan C Weller, and William H Batchelder. Culture as consensus: A theory of culture and informant accuracy. *American anthropologist*, 88(2): 313–338, 1986.

Lee Ross, David Greene, and Pamela House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301, 1977.

Peter E Rossi, Greg M Allenby, and Robert E McCulloch. *Bayesian statistics and marketing*. Wiley New York, 2005.

David S Scharfstein and Jeremy C Stein. Herd behavior and investment. *The American Economic Review*, pages 465–479, 1990.

Nicholas Seybert and Robert Bloomfield. Contagion of wishful thinking in markets. *Management Science*, 55(5):738–751, 2009.

Joseph P Simmons, Leif D Nelson, Jeff Galak, and Shane Frederick. Intuitive biases in choice versus estimation: implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15, 2011.

Mark Steyvers, Brent Miller, Pernille Hemmer, and Michael D Lee. The wisdom of crowds in the recollection of order information. In *Advances in neural information processing systems*, pages 1785–1793, 2009.

Jerry Suls and Choi K Wan. In search of the false-uniqueness phenomenon: fear and estimates of social consensus. *Journal of Personality and Social Psychology*, 52(1): 211, 1987.

Shyam Sunder. Market for information: Experimental evidence. *Econometrica: Journal of the Econometric Society*, pages 667–695, 1992.

Cass R Sunstein. The law of group polarization. *Journal of political philosophy*, 10 (2):175–195, 2002.

Cass R Sunstein. *Infotopia: How many minds produce knowledge*. Oxford University Press, USA, 2006.

James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

Phillip E Tetlock. *Expert political judgment: How good is it? How can we know?* Princeton University Press, 2005.

Donald N Thompson. *Oracles: How Prediction Markets Turn Employees into Visionaries*. Harvard Business Press, 2012.

R Scott Tindale and Katharina Kluwe. Decision making in groups and organizations. *The Wiley Blackwell Handbook of Judgment and Decision Making*, pages 849–874, 2015.

Maggie E Toplak, Richard F West, and Keith E Stanovich. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7):1275–1289, 2011.

Brandon M Turner, Mark Steyvers, Edgar C Merkle, David V Budescu, and Thomas S Wallsten. Forecast aggregation via recalibration. *Machine Learning*, 95(3):261–289, 2014.

Martin Waitz and Andreas Mild. Corporate prediction markets: a tool for predicting market shares. *Journal of Business Economics*, 83(3):193–212, 2013.

Peter Welinder, Steve Branson, Serge J Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *NIPS*, volume 23, pages 2424–2432, 2010.

Susan C Weller. Cultural consensus theory: Applications and frequently asked questions. *Field methods*, 19(4):339–368, 2007.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

Robert L Winkler. The consensus of subjective probability distributions. *Management Science*, 15(2):B–61, 1968.

Robert L Winkler. Combining probability distributions from dependent information sources. *Management Science*, 27(4):479–488, 1981.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

Yan Yan, Glenn M Fung, Rómer Rosales, and Jennifer G Dy. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168, 2011.

Sheng Kung Michael Yi, Mark Steyvers, Michael D Lee, and Matthew Dry. Wisdom of the crowds in traveling salesman problems. *Memory & Cognition*, 39:914–92, 2010a.

Sheng Kung Michael Yi, Mark Steyvers, Michael D. Lee, and Matthew J Dry. Wisdom of the crowds in minimum spanning tree problems. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010b.

Sheng Kung Michael Yi, Mark Steyvers, Michael D Lee, and Matthew J Dry. The wisdom of the crowd in combinatorial problems. *Cognitive science*, 2012.

Hang Zhang and Laurence T Maloney. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6:1, 2012.