

**Making Crowds Wiser: the role of  
incentives, individual biases, and improved  
aggregation**

Benjamin Tereick

ISBN: 978-90-361-0627-6

Cover illustration by Sieglinde Sobkowiak.

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul.

This book is no. **774** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

**Making Crowds Wiser: the role of incentives,  
individual biases, and improved aggregation**

**Crowds wijzer maken: de rol van incentives,  
individuele biases en verbeterde aggregatie**

Thesis

to obtain the degree of Doctor from the Erasmus University Rotterdam  
by command of the rector magnificus

Prof.dr. F.A. van der Duijn Schouten

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on 04 February, 2021 at 10:30 AM by

**Benjamin Tereick**

born in Duisburg, Germany

## Doctoral Committee

### *Promotors*

prof.dr. A. Baillon

prof.dr. D. Prelec

### *Other members*

prof.dr. K.I.M. Rohde

prof.dr. T. Offerman

prof.dr. S. Trautmann

# Preface

A guiding idea underlying this thesis is that when making decisions, it is often advisable to take the insights of many different people into account. I have diligently followed my own advice in this regard, as this thesis has profited from the ideas, suggestions and support of a great deal of people. In good decision-theoretic tradition, they will be named in backward order (taking some liberties), starting from those who contributed to the completion of this dissertation, and solving backwards to its origin.

Positioned at the final end of this chain are my supervisors. I am thankful for Aurélien Baillon to have advised me on just about every aspect of academic life. From research – ranging from the discussion of big picture questions to correcting the indexing in appendix proofs – to dealing with students, co-authors, journals and presentations, he not only continuously helped me to improve, but also led as a glowing example.

He was also the one to put me in contact with the second supervisor of this thesis, Dražen Prelec. It was a great honour to be able to discuss many of the methods in this thesis with the person who created an entire research field devoted to dealing with unverifiable subjective data – a topic at the heart of three chapters in this thesis. With his original and refreshing perspective on just about every topic, meetings with Dražen were always a delight and more than sometimes tremendously improved my understanding. It is said that “Poor is the pupil who does not surpass his master(s)”. If that is true, I expect to stay a rather poor pupil.

During the last four years, I have spent many hours with my office mate Yan Xu,

discussing not only research, but various topics ranging from personal life to the intricacies of the Chinese education system and German case reflections (good luck in Vienna, Yan!). I have had the pleasure to share a PhD office with a marvelous view over the Maas river with Merel van Hulsen, Francesco Capozza, Cem Peker and Xiao Yu. The Corona outbreak shortened this enjoyable time to just one month – but I already miss it! I will also remember the afternoons I spent with Tong Wang over buggy MATLAB code and evasive mathematical theorems, the result of which can be seen in Chapter 3 of this thesis. I have generally been glad to be a member of the behavioral economics group at Erasmus University. It has been great to be a colleague of Han Bleichrodt, Paul van Bruggen Aysil Emirmahmutoglu, Đura-Georg Granić, Jan Heufer, Chen Li, Kirsten Rohde, Vitalie Spinu, Jan Stoop, Jingni Yang and Sophie van der Zee. It is great that in this group, junior members are not regarded as cheap sources of labour, but trained as the researchers of the next generation.

This dissertation also marks the end of my time at the Tinbergen Institute, which started when I entered the Tinbergen research master in 2014. I want to thank Maria Cotofan, Sam Taylor, David Garces Urzainqui, Pim Kastelein, Laura Remie, Timo Klein, Magda Rola-Janicka, Robin Döttling and Huaiping Yuan for Sinterklaas presents, King’s Day celebrations, poker nights, fancy and not-so-fancy dinners, and so much more. I also don’t want to miss the time I spent with Jenny Kudymowa, Vadim Nelidov, Huyen Nguyen, Sanna Stephan and Dieter Wang during the early days in Amsterdam which helped me get through the bleak Dutch winter.

I am also grateful for the help of the Tinbergen Institute, which started with correspondence with an understanding Judith van Kronenburg over an overdue English proficiency test at my application, continued with Ester van den Bragt listening to a near-nervous-breakdown at the beginning of the MPhil and Christina Månsson and Carine Horbach handling my job applications in Fall 2019, and found its logical conclusion with Carolien Stolting guiding me through the printing process of this thesis.

From before my Dutch times, I would like to give a special thanks to Bernhard Stanke-witz, for believing in my talents and – I think unconsciously – pushing me to start believing in them again as well. I also want to give a collective thank you to the community of effective altruism, of which I became aware during my graduate studies and which has given me the opportunity to find people who share my eccentric view of the world. I am glad to have now found an academic position which allows me to stay in close contact with this very community.

I credit my family for instilling in me an (imperfect) sense of compassion and a desire for seeking the truth. Therefore, thanks to my mother and her husband Wolfgang, to my father, and to my sisters Jana and Miriam and their spouses Carl Friedrich and Baptiste. And finally, Sarah, I want to thank you for wandering the earth with me for the last eleven years. I hope there are many more to come – I can barely wait.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Market Scoring Rules</b>	<b>7</b>
2.1	Introduction . . . . .	7
2.2	Market Scoring Rules Illustrated . . . . .	11
2.3	The Model – Information Structure . . . . .	13
2.4	The Model – Strategic Behavior . . . . .	20
2.5	The Logarithmic Market Scoring Rule Game - Results . . . . .	27
2.6	Discussion . . . . .	34
2.7	Concluding Remarks . . . . .	39
<b>3</b>	<b>Follow the money</b>	<b>41</b>
3.1	Introduction . . . . .	41
3.2	Theory . . . . .	44
3.3	Experimental design . . . . .	58
3.4	Results . . . . .	62
3.5	Discussion . . . . .	70
3.6	Conclusion . . . . .	74
3.7	Proofs . . . . .	75
3.8	Proof of the lemmas . . . . .	82

3.9	Experimental instructions and training rounds . . . . .	93
<b>4</b>	<b>Honesty via Choice-Matching</b>	<b>95</b>
4.1	Introduction . . . . .	95
4.2	Example . . . . .	96
4.3	Model and Results . . . . .	98
4.4	Choice-Matching Generalized . . . . .	106
4.5	Comparison with Existing Methods . . . . .	110
4.6	Conclusion . . . . .	111
4.7	Additional Material . . . . .	113
<b>5</b>	<b>Aggregation Through Meta-Cognition</b>	<b>127</b>
5.1	Introduction . . . . .	127
5.2	Illustration: A Jury Decision . . . . .	131
5.3	Model . . . . .	134
5.4	Experimental Test . . . . .	154
5.5	Empirical Results . . . . .	160
5.6	Discussion . . . . .	167
5.7	Conclusion . . . . .	171
5.8	Proofs for Section 5.3 . . . . .	172
5.9	Task Interface for CONF, PRED and THRESH . . . . .	178
<b>6</b>	<b>Conclusion</b>	<b>181</b>
	<b>Summary</b>	<b>185</b>
	<b>Samenvatting</b>	<b>187</b>
	<b>Bibliography</b>	<b>206</b>

# Chapter 1

## Introduction

This thesis is about methods to receive information from groups of individuals – possibly very large groups – and to map this information into a decision. I have spent the last four years working on such methods because I am convinced that for many high-stake decisions, relevant bits of knowledge are held by a large number of people and that there is a lot of room for improvement in how we utilize this knowledge. Throughout the thesis, I focus on methods which reward or aggregate judgments in *probabilistic* form. The underlying framework in doing so is Bayesian decision theory, both in its descriptive and in its normative form. Descriptively, Bayesian decision theory is used to describe the knowledge possessed by an individual as a probability distribution over possible worlds. Normatively, Bayesian decision theory is used to evaluate which methods of combining knowledge is best.

All methods which try to induce large numbers of people to contribute to the solution of a problem face numerous challenges, and the probabilistic judgment aggregation and reward systems analyzed in this thesis are no exception. Two particularly important challenges are what will be called the “incentive problem” and the “aggregation problem”. The incentive problem is the problem of choosing rewards for information revelation such that all members of the human ‘crowd’ have reason to seek information, and will accurately

report it in the process. The aggregation problem concerns how to map the individual judgments of all crowd members into a single decision.

All four chapters in this thesis are devoted to at least one of these problems. Moreover, the methods described in these chapters aim to be simple enough such that humans, with all our cognitive shortcomings, can interact with them – not only highly mathematically sophisticated, dispassionate Bayesians.

## Chapter 2

In this chapter, I analyze the strategic aspects of the logarithmic market scoring rule (LMSR) – a betting system which was proposed by Robin Hanson (2003; 2007). The important gap in the literature filled by the LMSR is that Bayesian decision-makers usually should not be willing to bet against each other: if somebody else disagrees with your probabilistic assessment, this should be reason to revise your assessment, not to bet against them. In the LMSR however, people never bet directly against each other. Instead, the designer of the betting market – the “market-maker” – makes a first probabilistic statement about whether some potential event in the future will occur. For instance, by 2035, there may have been (i) another Moon landing (ii) the first human Mars mission, (iii) both a Moon and a Mars landing (iv) neither of the two. The market-maker then sets a probability for each of the possibilities (i)-(iv). Anyone can sign up to the market and make another probabilistic judgment about the future. They will be rewarded for the accuracy of this judgment, but in exchange must commit to rewarding the market-maker for the initial statement. Any subsequent market participant is also ensured to be rewarded, but in return must pay the reward for the previous participant.

The LMSR allows information to flow into the market continuously, until the future event resolves. I show in Chapter 2, that this scheme is also robust against manipulations, in the sense that truthful and immediate revelation of information is an equilibrium. That

is, if other respondents expect everyone else to update their probabilities correctly and to be truthful – then it is in their best interest to do and be so as well. This extends an earlier result by Chen et al. (2010). In their model, the betting market is on a single proposition which can be true or false. The question on space missions could then only take simpler forms, e.g. “Will there be a successful Mars mission by 2035 or not?”. In contrast, my result can be applied to possibilities such as (i)-(iv) above and in my model, participants can repeatedly receive new information. In the example, they could then, for instance, react to announcements by NASA, ESA or SpaceX, or on their private conversations with aerospace engineers. In addition to the formal treatment of these two generalizations, I discuss other generalizations informally, and make various suggestions for future work.

## Chapter 3

This chapter is based on the paper “Follow the Money: Bayesian Markets to Extract Crowd Wisdom” which is on-going joint work with Aurélien Baillon and Tong V. Wang. In contrast to Chapter 2, we are interested in beliefs about events which are not *verifiable*, building on pioneering work by Prelec (2004).

In prediction markets such as the LMSR, probabilistic statements are rewarded by comparing them to the true events as they unfold. For some statements, this is impossible. This is so, for instance, for answers to questions which relate to subjective experience. There is no reliable way to verify whether you are responding truthfully to the question “How happy are you right now?”. Verifiability also fails for questions relating to the very far future. Suppose for example that the question on space missions is not on a Mars or moon mission until 2035, but whether humanity will spread to other star systems before the year 2200. Yet another question with unverifiable answers would be a question about counterfactual worlds, such as “How would world history have unfolded, if the Mongol empire had persisted?”.

Just as the LMSR considered in Chapter 2 simultaneously solves the incentive problem and the aggregation problem for forecasts of verifiable events, the “Follow the money” (FTM) mechanism studied in Chapter 3 simultaneously solves the incentive problem and the aggregation problem when verification is impossible. We ensure incentives by using “Bayesian markets” introduced by Baillon (2017) and the underlying method of aggregation is based on the Surprisingly Popular Algorithm (SPA) introduced by Prelec et al. (2017). Building on a similar theoretical setting as those two papers, my co-authors and I show that for sufficiently large crowds, the SPA optimally aggregates information. Furthermore, selecting an answer according to those respondents who made the most money on the Bayesian market (thereby “following the money”) approximates the selection of the SPA – and is thus itself (approximately) an optimal aggregator. This is encouraging as Bayesian markets elicit responses by simple economic means, with simple, binary bets which require limited cognitive effort from each individual crowd member. This makes them an attractive method for practical implementation. Majority voting – arguably the most natural form of judgment aggregation – will often fail to make optimal use of the information provided by the crowd, even when the crowd is large.

We further suggest empirical methods to enrich the application of the FTM in practice, making it more functional for decision-makers who do to not fully adhere to Bayesian updating. A limitation of the FTM is that it is restricted to a setting in which the underlying question we want to be answered can only have two possible answers.

We test the FTM in a large online sample of US students. We used a task developed for the experiment reported in Chapter 5 that ensures that the informational assumptions of the model are satisfied. We compared our method to the majority opinion and to the SPA. Despite using less information than the SPA, our method had comparable accuracy rates. Both the FTM and the SPA substantially improved upon majority voting.

## Chapter 4

Just like Chapter 3, Chapter 4 deals with beliefs about events which are not verifiable, extending the incentive scheme discussed in Chapter 3. This chapter is based on work I published with Jakša Cvitanić, Dražen Prelec and Blake Riley. We introduce choice-matching, an incentive compatible mechanism for eliciting honest responses to a multiple choice question (MCQ), such as the space mission question discussed above – allowing any number of possible answers.

Under choice-matching, respondents are compensated for an auxiliary task, e.g., a personal consumption choice or a forecast. Their compensation depends both on their performance on the auxiliary task, and on the performance of those respondents who matched their response to the MCQ.

The idea is to link explicit opinions or judgments, which are answers to an MCQ, with an auxiliary task that reveals the respondents’ “types”, but only implicitly. A guiding example we use is one of a firm which launches a new product, and supplies a panel of potential customers with a trial product. How much each panel member liked the trial product is an unverifiable question, of the type relating to subjective experience.

To see how choice-matching works in this example, suppose that the firm asks each member to rate the hypothetical product on a scale from 1 to 5 stars and to additionally give a monetary evaluation for a product which already exists. The firm can use an incentive compatible pricing mechanism to incentivize the latter. The payoff of this task can then be linked to the answer to the MCQ: With some positive probability, the firm uses the respondents’ own reservation price for the price mechanism and with the remaining probability uses the average reservation price of those respondents who gave the same rating to the hypothetical product. When the valuations of the trial product and the existing product are sufficiently linked, this scheme implements honest valuations in a strict Bayes-Nash Equilibrium.

We further show how to create an auxiliary task when a plausible link like in the example above is not obvious. In this “canonical version” of choice-matching, the auxiliary task asks respondents to predict how often each answer option in the MCQ was chosen by the other respondents. A respondent’s score is then a weighted sum of a prediction accuracy score and the average prediction accuracy score of all the respondents who endorse the same answer to the MCQ. We show that under plausible assumptions about the updating process, truthfully revealing the answer to the MCQ and predictions is again a strict Bayes-Nash Equilibrium.

## Chapter 5

The final chapter zooms in on the aggregation problem. I show that the Surprisingly Popular Algorithm proposed by Prelec et al. (2017) – and therefore the scheme analyzed in Chapter 3 – is not an optimal aggregator for Bayesian decision-makers if the sample is small.

To overcome this issue, I propose “self-aggregation” (SELF), a new aggregation scheme. In a binary choice setting, SELF asks individuals to vote for an option and to simultaneously provide a threshold of the number of people that would convince them of the opposite. SELF then picks an option if more people vote for it than the average threshold provided in the group. In a model in which individuals update their beliefs in a Bayesian fashion, SELF has a higher probability of leading to the correct decision than a number of benchmarks; other than the SPA, I consider majority voting and different methods based on confidence-levels. In an experimental test of the model, respondents solve a binary decision problem in a stylized urn experiment. Responses and aggregation results can be directly compared to the Bayesian prescription. In the experiment, SELF compares favorably to confidence-based methods, but does not realize its theoretical advantage over the SPA.



# Chapter 2

## Market Scoring Rules in a Dynamic Bayesian Setting

### 2.1 Introduction

Hanson (2003) proposed the use of *market scoring rules*. A market scoring rule (MSR) is a subsidized betting market in which agents sequentially report beliefs about a future event and are rewarded for their accuracy. A market maker sets a baseline probability report and subsequent market participants report their own belief. In doing so, they agree to insure the market-maker against the previous report, in exchange for being rewarded for their own. As a result, the market-maker only ever pays for one report. In this chapter, I show that a specific market scoring rule – the *logarithmic* market scoring rule (LMSR) – induces participants to report their knowledge truthfully and instantaneously, and aggregates the information held by all participants for a fairly large class of information structures. A crucial underlying assumption is that individual information is independent conditional on the outcome of the future event. This means that what two different individuals believe is only informative about each other because both are related to the actual outcome.

The result proved in this chapter further strengthens market scoring rules as a candi-

date to solve a puzzle at the heart of economic theory: namely to provide a rational foundation of an institution which exchanges individual bets. Starting with Ramsey (1931), De Finetti (1931) and Savage (1954), subjective probabilities – a key ingredient of economic decision theory – have been interpreted as a “willingness to bet”.<sup>1</sup> This acceptance to bet – or, analogously, to sell or buy a contingent contract – is behind classic economic results on the information properties of markets. And indeed, in the real world, we can observe betting as a widespread social phenomenon. Bets are placed on a diverse spectrum of events, ranging from the geo-political sphere to celebrity gossip and sports. The theoretical challenge however in economic theory is that the micro-level specifics of the macro-level information property are unclear: When two decision-makers disagree on the probability of a future event, neither should accept to bet on it, as not both of them can rationally expect to make a profit from this bet.<sup>2</sup>

This negative result may explain why despite promising practical trials (e.g. Forsythe et al., 1992; Gillen et al., 2017) and a generally supportive attitude among economists (Arrow et al., 2008, Wolfers et al., 2012), prediction markets have not yet played a central role in societal decision making – or as an object of economic study.

Hopefully, this chapter adds to our theoretical understanding of how to design prediction markets and can further promote their study and practical implementation.

### 2.1.1 Related Literature

Market scoring rules were proposed by Hanson (2003) and a particular case for the use of the logarithmic market scoring rule (to be defined in Section 2.4) was made in Hanson (2007). Hanson discusses consistency with Bayesian updating only informally, focusing

---

<sup>1</sup>For instance, if a decision-maker believes that event  $E$  occurs with probability  $p$  and that event  $E'$  occurs with probability  $q < p$ , then the decision-maker should prefer to receive some prize when  $E$  occurs, rather than on the condition that  $E'$  occurs. Similarly, when choosing to receive or pay the same amount if  $E$  does or does not occur, the decision-maker should prefer to bet on the event  $E$  if and only if it is more likely than not to occur (or  $p > \frac{1}{2}$ ).

<sup>2</sup>The canonical articulation of this idea is the no-trade theorem of Milgrom and Stokey (1982).

on implementation of market scoring rules for event spaces in which the market-maker is interested in finding efficient ways to allow betting on conditional beliefs. Furthermore, while Hanson discusses the model in terms of a time dimension, this dimension is not explicit in the model.

A survey of different forms of prediction markets, including market scoring rules, is Chen and Pennock (2010). As pointed out already by Hanson (2003), and analyzed more thoroughly by Chen and Pennock (2007), any market scoring rule has an equivalent formulation as a more traditional prediction market in which individuals can buy Arrow securities (which pay 1 unit of currency if an event occurs, and nothing otherwise). Upon each individual purchase of a security, prices are adapted using a pricing mechanism which makes the decision problem of a market participant equivalent to a market scoring rule. The properties of market scoring rules ensure that individual buying decisions can be interpreted as individual probability judgments, avoiding problems with the interpretation of prices in traditional, pari-mutuel prediction markets pointed out by Manski (2006 – see also Wolfers and Zitzewitz, 2006). Chen and Vaughan (2010) relate market scoring rules to no-regret learning.

The Arrow security formulation of market scoring rules is used in applied work by Dreber et al. (2015) and Camerer et al. (2016) and in on-going follow-up work. According to Chen and Pennock (2010), it has also been used for internal analytics by Inkling Markets, Consensus Point, Yahoo!, and Microsoft. Translating market scoring rules to a pricing mechanism has the advantage of offering a more familiar environment to respondents. While the pricing rule may be opaque to some respondents, each purchasing decision is a simple trade of a security for a single price.

Ostrovsky (2012) discusses market scoring rules in the context of the general information aggregation properties of markets and allows only *myopic* strategic behavior, i.e. behavior in which participants do not condition their beliefs on the beliefs of others – and don't expect others to do so. When participants are myopic, it is straightforward that

MSRs induce truthful information revelation. Accordingly, Ostrovsky uses market scoring rules to illustrate information aggregation without the problem of revelation. Similarly, Carvalho (2017) studies aggregation properties of equilibrium prices under the logarithmic market scoring rule for myopic traders. Strategies are also assumed to be myopic in Dimitrov et al. (2015) who consider risk-averse traders. Buckley and O’Brien (2017) experimentally test the effect of manipulations in a real prediction market. They show that manipulations of market prices reduce the accuracy of prediction market forecast in the short-term, but are corrected by other traders relatively quickly. While this is an encouraging finding, fear of manipulation may still lead traders to choose not to participate, or not to fully reveal information in the market.

Closest to the model and results in this chapter is Chen et al. (2010), which combines results by Chen et al. (2007) and Dimitrov and Sami (2008). Chen et al. (2010) study the robustness of the logarithmic market scoring rule to manipulation for a market which trades forecasts on a binary event, and in which each market participant observes a single signal as the market opens. In contrast, in the model I introduce below, market participants continuously receive information and the event can be non-binary. Especially allowing more than a single period in which market participants receive private information is an important contribution of this chapter. While the model in Chen et al. (2010) is dynamic in the sense that participants are allowed to change their forecasts over time, there is little reason for doing so since truth-telling would imply that all participants report their forecast in the first period. The strategic aspects analyzed in their paper could then be solved by simply letting every market participant report once, and only once. The repeated updating thus only seems sensible if we assume that information can arrive at various times, which indeed seems like a more accurate description of prediction markets in the real world.

The novel dynamic element of the model in this chapter makes the setting similar to the dynamic belief elicitation problem discussed in Chambers and Lambert (2020) who only

consider single decision-makers, but allow that decision-makers may not be risk-neutral or have preferences over the forecasted outcomes. Future work on prediction markets could likely profit from synergies with this related dynamic elicitation problem.

## 2.1.2 Organization of this Chapter

The rest of this chapter is organized as follows: In Section 2.2, I provide an example of how market scoring rules work. Sections 2.3 and 2.4 constitute the formal model, where Section 2.3 introduces the information structure underlying the market and Section 2.4 describes the incentives provided by market scoring rules as an extensive form game. The main result that the restricted logarithmic market scoring rule (LMSR) induces honest information revelation is proven in Section 2.5. Other practical aspects and properties of the LMSR are discussed informally in Section 2.6. These issues are fertile ground for future formal analysis. I conclude in Section 2.7.

## 2.2 Market Scoring Rules Illustrated

Suppose that we wish to forecast the GDP growth of India in the following year. For simplicity, suppose we are just interested in the probability that GDP growth will be above or below 5%. Suppose further that we would like to get a variety of opinions, and thus do not like to ask only a specific set of experts for an opinion, but rather would like to crowd-source the forecast. How could people be encouraged to participate in such a forecasting exercise, and how can they be encouraged to provide forecasts to the best of their knowledge? One way to do so, would be to reward each single participant with a *proper scoring rule*.

The idea behind a proper scoring rule is as follow: A forecaster  $i$  reports a probability  $r_i$  that GDP growth will be above 5%. Let  $E = 1$  denote that GDP growth will be above 5% and  $E = 0$  denote that GDP growth will be below 5%. After observing the actual

GDP in the following year, we can assign a score to forecaster  $i$  according to a scoring rule  $R(r_i, E)$ . Suppose then that the forecaster  $i$ 's subjective belief that GDP growth will be above 5% is  $p_i$ . The forecaster can then calculate  $p_i R(r_i, 1) + (1 - p_i) R(r_i, 0)$ , the subjectively expected score for making report  $r_i$ . A proper scoring rule  $R$  has the property that maximizing this expectation with respect to  $r_i$  yields  $r_i = p_i$  as the unique solution.<sup>3</sup> From the perspective of the forecaster, it is thus optimal to report the true belief  $p_i$ .

The problem with using a single scoring rule for each participant is that it could be very costly. Letting the forecasters bet against each other could reduce those costs.<sup>4</sup> However, if the forecasters are sophisticated Bayesian agents, they will not accept bets against each other, as they should immediately agree about the GDP forecast upon learning about their disagreement (Aumann, 1976, Milgrom and Stokey, 1982).

Hanson (2003) proposed a solution. Suppose there are  $n$  participants in the crowd-sourced forecasting project and let  $p_1, \dots, p_n$  represent each forecaster's probabilistic belief that  $E = 1$ . The market-maker sets an initial "probability report"  $r_0$ . Afterwards, the first forecaster can change this probability report to  $r_1$ , the second forecaster can change the report to  $r_2$ , and so on. After every forecaster has made their forecast, and GDP growth materializes, the  $i$ -th forecaster receives a score  $R(r_i, E) - R(r_{i-1}, E)$ , where  $R$  is a proper scoring rule. When  $R$  is the logarithmic scoring rule (given by  $R(r_i, 1) = \log(r_i)$ ,  $R(r_i, 0) = \log(1 - r_i)$ ), the market scoring rule is called the *logarithmic market scoring rule* (LMSR).

Since  $R$  is proper and the  $i$ -th forecaster cannot influence the score  $R(r_{i-1}, E)$ , the score  $R(r_i, E) - R(r_{i-1}, E)$  also constitutes a proper scoring rule. Moreover, propriety of  $R$  ensures that each forecaster expects their score to be positive (since we must have that

---

<sup>3</sup>Scoring rules with this property are often called "strictly proper scoring rule", with a proper scoring rule not requiring uniqueness. Since non-strictly proper scoring rules are not particularly interesting in the context of this chapter, I use the simpler term here.

<sup>4</sup>As will become clearer after the introduction of the model, the cost difference for individual elicitation and elicitation via a betting market is essentially that for individual elicitation, a market-maker will have to pay for shared information, whereas in a betting market, each participant receives the marginal contribution of their information in expectation.

$i$  expect  $R(p_{i-1}, E)$  to be smaller than  $R(p_i, E)$ ). This way of incentivizing thus ensures both participation and truthful reporting. Moreover, note that because the forecasters mutually insure each other, the market-maker's cost is just  $R(r_n, E) - R(r_0, E)$ . In particular, the maximal cost does thus not change with the number of forecaster – a crowd of 100,000 forecasters could be financed by the same funds as a panel of five.

Despite its elegance, Hanson's argument leaves some open questions. First, we have fixed the forecasters' beliefs  $p_1, \dots, p_n$ . It would be reasonable to expect them to include the information provided by previous reports. In fact, it seems desirable from the market-makers perspective that forecasters do so, such that the final report includes all information held by participants. However, if market participants update their beliefs upon learning each others' reports, is there value in misleading others about one's information? Similarly, could there be value in delaying entry – in order to profit from the information provided by other participants?

As I show in this chapter, the logarithmic market scoring rule is robust to manipulation and encourages early entry under fairly general conditions. This is so even if respondents repeatedly receive information until the forecasted event materializes, and are allowed to make multiple forecasts. The next two sections introduce the terminology and assumptions used in order to formally state and prove this result.

## 2.3 The Model – Information Structure

In the next two sections, I will formally describe the prediction market defined by a market scoring rule as a game between its participants. I start with describing the information structure underlying the market. Strategic aspects are modeled in Section 2.4.

### 2.3.1 The Information Structure - Setup and Notation

There is a set of time periods  $\mathcal{T} = \{1, \dots, T\}$  and a set of individuals  $N = \{1, \dots, n\}$  – the individuals who can participate in the market. In time period  $T+1$ , the future event  $e \in \mathcal{E}$  materializes, where  $\mathcal{E} = \{1, \dots, n_E\}$ . We would like to elicit beliefs from individuals for each of the  $n_E$  possible outcomes of the event. In each time-period  $t \in \mathcal{T}$ , individuals receive a *signal*  $s_t^i \in \mathcal{S} = \{0, 1, \dots, n_S\}$ . Signals with value zero are *uninformative* signals in a sense defined below. The vector  $h_t^i = (s_1^i, \dots, s_t^i)$  is *i's signal history at t*.  $h^i = h_T^i$  is *i's signal history* (without qualification), and  $h = (h^1, \dots, h^n)$  is a *signal history* (for all the players). If there is no ambiguity, I will also call  $h_t^i$  a signal history, or just a history.  $\mathcal{H}$  and  $\mathcal{H}_t$  (for  $t \in \mathcal{T}$ ) denote the set of all possible signal histories and of all signal histories at  $t$ , respectively. It will also sometimes be convenient to allow for a history at time 0, with  $\mathcal{H}_0$  only containing the empty set. I take  $h_t^{-i}$  to be the collection of all histories of individuals other than  $i$ , with the twist that it contains  $j$ 's history at  $t$  if  $j < i$  and  $j$ 's history at  $t-1$  if  $j > i$ .<sup>5</sup>

Together with the set of all the possible values of  $e$ , the set of all possible signal histories forms the state space  $\Omega = \mathcal{E} \times \mathcal{H}$ . I denote by  $\mathbb{P}$  a probability measure on the sigma algebra  $\mathcal{F} = 2^\Omega$ . As will be formally captured through the equilibrium concept in Section 2.4, the probability measure  $\mathbb{P}$  is common knowledge among all respondents, but signals are only privately known. With respect to the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , each individual  $i$ 's signal at time  $t$  is a random variable  $S_t^i$ , so that each  $\omega \in \Omega$  corresponds to a specific event  $S_t^i = s_t^i$ , where  $s_t^i \in \mathcal{S}$  will be called a “realization” of  $S_t^i$ . I will use capital letters to denote random variables and tuples of random variables throughout, so that we have the events  $E = e$ ,  $H^i \equiv (H_1^i, \dots, H_T^i) = h^i \equiv (h_1^i, \dots, h_T^i)$ , and so on.

---

<sup>5</sup>This small complication in definition will be useful due to the timing provided by market scoring rules: When  $i$  is to move at  $t$ ,  $i$  knows the reports of all individuals  $j < i$  from period  $t$ , but for  $j > i$ , only the reports up until period  $t-1$ . While the timing structure of the model assumes that signals of all individuals at  $t$  are determined simultaneously, the notation  $h_t^{-i}$  will be mostly used in the strategic analysis of market scoring rules, and thus defining it according to the market scoring rule timing is useful.



With slight abuse of notation, for an event  $f \in \mathcal{F}$ , I write  $\mathbb{P}(E|f)$  for the vector of probabilities  $(\mathbb{P}(E = 1|f), \dots, \mathbb{P}(E = n_E|f))$  and analogously  $\mathbb{P}(f|E)$  for the vector of probabilities  $(\mathbb{P}(f|E = 1), \dots, \mathbb{P}(f|E = n_E))$ . I further write  $p_0 = \mathbb{P}(E)$  for the common belief about the event at the beginning of the market.

### 2.3.2 The Information Structure - Properties

I will next define some assumptions on  $\mathbb{P}$  which are made throughout this chapter.

**Definition 1.** Two events  $f, g \in \mathcal{F}$  are *conditionally independent with respect to  $E$*  if we have :

$$\mathbb{P}(f|g, E) = \mathbb{P}(f|E).$$

Furthermore, two random variables  $X$  and  $Y$  on  $(\Omega, \mathcal{F}, \mathbb{P})$  are conditionally independent with respect to  $E$  if the events  $X = x$  and  $Y = y$  are conditionally independent with respect to  $E$  for all realizations  $x, y$ .

Next, let  $f, g \in \mathcal{F}$  be two events which are conditionally independent with respect to  $E$  and let  $\mathbb{P}(E|f)$  reflect the “current state of information” after learning about  $f$ . Then, learning in addition that  $g$ , we get for all  $e \in \mathcal{E}$ :

$$\begin{aligned} \mathbb{P}(E = e|f, g) &= \mathbb{P}(E = e|f) \frac{\mathbb{P}(g|E=e)}{\mathbb{P}(g|f)} \\ &= \mathbb{P}(E = e|f) \frac{\mathbb{P}(g|E=e)}{\sum_{e' \in \mathcal{E}} \mathbb{P}(E=e'|f) \mathbb{P}(g|E=e')} \end{aligned}$$

And in more compact notation,

$$\mathbb{P}(E = e|f, g) = \frac{\mathbb{P}(E = e|f) \mathbb{P}(g|E = e)}{\langle \mathbb{P}(E|f), \mathbb{P}(g|E) \rangle} \quad (2.3.1)$$

with  $\langle u, v \rangle$  denoting the dot-product of two vectors  $u$  and  $v$ . I will next relate the definitions above to the individual signals.

**Assumption 1.** For any pairs  $(i, t) \neq (j, u)$  where  $i, j \in N$  and  $t, u \in \mathcal{T}$ , the two random variables  $S_t^i$  and  $S_u^j$  are conditionally independent with respect to  $E$ .

Since this assumption permits that  $i = j$ , signals of the same individual at different time periods are also assumed to be conditionally independent. Note that conditional independence does not require that the signals of two different individuals, or of the same individual at different time periods, are identically distributed. It is also allowed that  $\mathbb{P}(S_t^i = s) = 0$  for some  $s \in \mathcal{S}$ .<sup>6</sup>

A second assumption on  $\mathbb{P}$  is stochastic relevance:

**Assumption 2.** For all  $i \in N$ ,  $t \in \mathcal{T}$ , and for all  $s, s' \in \mathcal{S}$ , the random variable  $S_t^i$  satisfies stochastic relevance (for  $E$ ). This means that for all  $s, s' \in \mathcal{S}$ :

$$\mathbb{P}(E | S_t^i = s) \neq \mathbb{P}(E | S_t^i = s') \quad (2.3.2)$$

Together, Assumptions 1 and 2 imply that for  $t \in \mathcal{T}$  and any  $h \in \mathcal{H}_t^{-i}$ :

$$\mathbb{P}(E | S_t^i = s, H_t^{-i} = h) \neq \mathbb{P}(E | S_t^i = s', H_t^{-i} = h) \quad (2.3.3)$$

Thus, stochastic relevance of individual signals persists after conditioning on the signals of other individuals. The relevance for the prediction market is that even after learning about the information held by others, each individual piece of information determines a unique probabilistic belief about the event  $E$ .

A further implication of Assumption 1 is that knowing  $\mathbb{P}(E | H_t^{-i} = h_t^{-i})$  and  $S_t^i$ , an individual  $i$  can use Equation 2.3.1 to calculate:

$$\mathbb{P}(E = e | S_t^i = s, H_t^{-i} = h) = \frac{\mathbb{P}(S_t^i = s | E = e) \mathbb{P}(E = e | H_t^{-i} = h)}{\langle \mathbb{P}(S_t^i = s | E), \mathbb{P}(E | H_t^{-i} = h) \rangle} \quad (2.3.4)$$

---

<sup>6</sup>As a result, this model nests the model in Chen et al. (2010), in which each individual can receive signals from a space  $\mathcal{S}_i$ . We can simply let  $\mathcal{S} = (\mathcal{S}_i)_{i \in N}$  and  $\mathbb{P}(S_i^t = s) = 0$  for all  $s \notin \mathcal{S}_i$ .

Thus, at time point  $t$ , an individual does not need to have access to the entire history  $h_{t-1}^{-i}$ . It is enough to know the probability vectors  $P(S_t = s | E)$  and  $\mathbb{P}(E = e | H_t^{-i} = h)$ . This reduces the computational burden for individuals.

The next assumption is that no signal history determines the event with certainty.

**Assumption 3.** *For all signal histories  $h \in \mathcal{H}$  and all  $e \in \mathcal{E}$ :*

$$\mathbb{P}(E = e | H = h) \neq 1 \tag{2.3.5}$$

This assumption is mostly made for technical convenience, but can also be seen as a normative commitment that no belief about an empirical event should be 1 until the event has realized. Finally, as mentioned above,  $S_t^i = 0$  denotes that  $i$  does not receive any new information at time  $t$ . We can now formally express this as follows: For all  $i \in I$  and all  $s, s' \in \mathcal{S}$ , for any  $t \in \mathcal{T}$  and any  $f \in \mathcal{F}$ :

$$\mathbb{P}(E | S_t^i = 0, f) = \mathbb{P}(E | f) \tag{2.3.6}$$

Note that this is a vacuous assumption (since we can always add 0 to any existing set of signals) unless we specify that  $\mathbb{P}(S_t^i) > 0$ . The purpose of including zero-signals is that we can capture that individuals do not receive information by describing them as receiving a specific signal. For the strategic setting, we can then treat individuals who are “waiting out” the market by claiming that their probabilistic beliefs have not changed as a special case of them misstating the information they have received. It is completely compatible with the model however, to make the assumption that individuals have to receive relevant information in every period.

### 2.3.3 Scoring Rules and KL-divergence

Before turning to the interactive aspects of strategic reports made by multiple decision-makers, it will be useful to recall the central concept of individual belief elicitation. A *scoring rule*  $R$  can be thought of as mapping a probabilistic forecast of  $E$  and the actual realization of  $E$  into a real-value:

$$R : \Delta^{n_E} \times \mathcal{E} \rightarrow \mathbb{R} \quad (2.3.7)$$

with  $\Delta^{n_E}$  the  $n_E$ -dimensional probability simplex. A *proper* scoring rule has the property that for any  $p \in \Delta^{n_E}$

$$\arg \max_{r \in \Delta^{n_E}} \sum_{e \in \mathcal{E}} p_e R(r, e) = p \quad (2.3.8)$$

This has the following interpretation, mirroring the explanation in Section 2.2: Suppose that a decision-maker observes event  $f \in \mathcal{F}$  and thus has subjective beliefs  $p = \mathbb{P}(E|f)$ . The decision-maker can then calculate the expected payoff  $\sum_{e \in \mathcal{E}} p_e R(r, e)$  when reporting to have the belief  $r \in \Delta^{n_E}$ . A proper scoring rule has the property that this expected payoff is maximized when  $r = p$ , i.e. when the decision-maker reports her true belief.

Closely connected to the notion of proper scoring rules is the measure of the *divergence* of two probability vectors  $x, y \in \Delta^m$ . The so-called *Kullback–Leibler divergence* (KL-divergence) is given by

$$L(x, y) = \sum_{k=1}^m x_k \log \left( \frac{x_k}{y_k} \right) \quad (2.3.9)$$

As can be easily checked, we have  $L(x, x) = 0$  and  $L(x, y) > 0$  whenever  $x \neq y$ , a fact known as the Gibbs' inequality.

To see the connection to proper scoring rules, note that we can restrict  $x$  to only take

values at the edges of the simplex  $\Delta^{n_E}$ . We could interpret these as outcomes of the event  $E$ , where  $x_e = 1$  denotes that  $E = e$ . Then, the scoring rule  $R(r, e) = c + \log r_e$  can be written as  $c - L(r, x)$ . This scoring rule is called the *logarithmic scoring rule* (LSR). Furthermore, when  $R$  is the logarithmic scoring rule, the maximization problem in 2.3.8 can be rewritten as minimizing  $L(p, r)$ . Since this minimization problem has the unique solution  $r = p$ , the LSR is proper.

A key feature of the information structure introduced thus far is captured in the following Lemma:

**Lemma 1.** *Let  $f, f' \in \mathcal{F}$  be two events and let  $\mathcal{G} \subset \mathcal{F}$  be a partition on  $\Omega$  such that any event  $g \in \mathcal{G}$  is, conditionally on  $E$ , independent of  $f$  and  $f'$ . Then, we have for any  $g \in \mathcal{G}$ :*

$$\frac{\mathbb{P}(E = e | f) \mathbb{P}(E = e | f', g)}{\mathbb{P}(E = e | f') \mathbb{P}(E = e | f, g)} = \frac{\mathbb{P}(g | f)}{\mathbb{P}(g | f')}$$

*Proof.* Using Bayes' rule:

$$\mathbb{P}(E = e | g, f') = \frac{\mathbb{P}(g | f', E = e)}{\mathbb{P}(g | f')} \mathbb{P}(E = e | f'),$$

$$\mathbb{P}(E = e | g, f) = \frac{\mathbb{P}(g | f, E = e)}{\mathbb{P}(g | f)} \mathbb{P}(E = e | f),$$

Since  $\mathbb{P}(g | f, E) = \mathbb{P}(g | f', E) = \mathbb{P}(g | E)$  by assumption, we get

$$\frac{\mathbb{P}(E = e | g, f)}{\mathbb{P}(E = e | g, f')} = \frac{\mathbb{P}(E = e | f) \mathbb{P}(g | f')}{\mathbb{P}(E = e | f') \mathbb{P}(g | f)}$$

which can be rearranged to yield the Lemma. □

We can derive the following from Lemma 1. Taking logs on both sides, we get:

$$\log \left( \frac{\mathbb{P}(E = e | f)}{\mathbb{P}(E = e | f')} \right) - \log \left( \frac{\mathbb{P}(E = e | f, g)}{\mathbb{P}(E = e | f', g)} \right) = \log \left( \frac{\mathbb{P}(g | f)}{\mathbb{P}(g | f')} \right).$$

The first term is the log-likelihood ratio for predicting the event  $E = e$  from events  $f$  and  $f'$ , the second term is the same ratio, but with additional conditioning on an event  $g$  which is conditionally independent of  $f$  and  $f'$ . The lemma says that the difference in these two log-likelihoods is the log-likelihood ratio for predicting the event  $g$  and thus does not depend on  $E$ . In particular thus:

$$\begin{aligned}
 & \sum_{g \in \mathcal{G}} \mathbb{P}(g|f) \left[ \log \left( \frac{\mathbb{P}(E=e|f)}{\mathbb{P}(E=e|f')} \right) - \log \left( \frac{\mathbb{P}(E=e|f,g)}{\mathbb{P}(E=e|f',g)} \right) \right] \\
 = & \sum_{g \in \mathcal{G}} \mathbb{P}(g|f) \log \left( \frac{\mathbb{P}(g|f')}{\mathbb{P}(g|f)} \right) \\
 = & L(\mathbb{P}(g|f), \mathbb{P}(g|f')) > 0.
 \end{aligned}$$

This property will be important when analyzing payment schemes building on the logarithmic scoring rule. Suppose that an individual predicts the event  $E = e$  and misstates their information as event  $f'$ , rather than  $f$ . Other individuals, conditioning on the misstated event  $f'$  also predict  $E$ , taking their own information  $g$  into account. The original individual, when taking expectations, according to the equation above, expects that the predictive error without conditioning on  $g$  is larger than with conditioning. Essentially, this property of an information structure in which individual private information is conditionally independent makes the LMSR robust to strategic manipulation. The details of this idea will be spelled out in the next section.

## 2.4 The Model – Strategic Behavior

In this section, we will treat market scoring rules with the tools of game theory. I first recall the definition of market scoring rules, in a slightly more general and formal way:

**Definition 2.** A market scoring rule is defined through the following procedure. Let  $\mathcal{R} = \text{int}(\Delta^{n_E})$ , the interior of the simplex  $\Delta^{n_E}$ . At period 0, the market maker sets the “initial belief”  $r_1^0$  equal to  $p_0$  (the prior probabilities regarding the event  $E$ ). At

every  $t \in \mathcal{T}$ , each individual  $i \in N$  makes a *report*  $r_t^i \in \mathcal{R}$ , upon learning all previous reports  $r_1^1, \dots, r_1^n, \dots, r_t^1, \dots, r_t^{i-1}$ . After all reports  $r = (r_1^1, \dots, r_T^n)$  are made, and the event  $E$  resolves, each individual  $i$  receives the payment:

$$\pi^i(r, e) = \sum_{t \in \mathcal{T}} R(r_t^i, e) - R(r_t^{i-1}, e), \quad (2.4.1)$$

where  $r_t^0$  is defined as  $r_{t-1}^n$  for all  $t > 1$  and  $R$  is a proper scoring rule. If  $R$  is the logarithmic scoring rule, then we speak of the *logarithmic market scoring rule* (LMSR).

Note that while the index ordering of individuals is deterministic in the model, one could in practice think of drawing them randomly before the start of the market, or even draw them anew at each period. See Section 2.6 for further discussion.

The definition is identical to the one in Hanson (2003) and in Chen et al. (2010), except for the restriction that each report has to be in  $\text{int}(\Delta^{nE})$ . This avoids the issue of infinite payoffs on off-equilibrium paths when the logarithmic scoring rule is used. As a further restriction, in what follows, I will only allow pure strategies for all players in  $N$ . This is purely to avoid notational complexities. Outcomes determined by mixed strategies would not be measurable by  $\mathbb{P}$ . With slight abuse of terminology, I will refer to an assessment in which each  $i \in N$  plays a pure strategy as a pure strategy profile. While the state of the world player  $W$  is in principle an “ordinary” player of the game, the fact that  $W$  has only one available strategy, makes it mute to specify  $W$ ’s strategy in a description of an assessment.

In principle, the payment scheme by an MSR defines an extensive form game of incomplete information, i.e. a game in which participants move consecutively and in which participants are uncertain about some aspects of the payment structure. This latter fact is since respondents do not know which future information they receive, what the outcome is of the event they are scored for, and so on. To be able to apply standard equilibrium concepts, a game of incomplete information is usually translated to a game of imperfect

information, in which all payments are specified but participants do not always know previous moves by other players. The player set then includes a “nature player” whose unique strategy is determining events according to the information structure (in case of the MSR, determining the event  $E$  and informing individuals about their signals  $S_t^i$ ).

### 2.4.1 Extensive Form Games

The following definition of an extensive form game is similar to the canonical one given in Mas-Collel et al. (1995) and Osborne and Rubinstein (1994).

**Definition 3.** A game  $\Gamma$  is a collection  $\langle \mathcal{P}, \mathcal{A}, \mathcal{D}, \succ, \mathcal{O}, o, u, \mathcal{I}, \alpha, \phi \rangle$ . The set  $\mathcal{P}$  is a finite set of players and  $\mathcal{A}$  is a set of actions.  $\mathcal{D}$  is a set of *decision nodes* and each  $d \in \mathcal{D}$  is a sequence of actions  $a_1, \dots, a_n, a_i \in \mathcal{A}$ .  $\succ$  is a successor relation on  $\mathcal{D}$ , where a node for which there is no successor is called a *terminal node*.  $\mathcal{O}$  is a set of outcomes and the *outcome function*  $o$  maps each terminal decision node into an outcome.  $(u_i)_{i \in \mathcal{P}}$  is a collection of utility functions which map  $\mathcal{O}$  into the real numbers.  $\mathcal{I}$  is a collection of information sets, forming a partition on  $\mathcal{D}$  (so that each information set in  $\mathcal{I}$  is a set of decision nodes), the *action function*  $\alpha$  assigns a set of available actions to each information set and the *player function*  $\phi$  assigns each information set to a player of the game.

In intuitive terms, an extensive form game of imperfect information is played by players  $\mathcal{P}$  who move sequentially at decision nodes  $d$ , with the player function  $p$  determining who moves at which point.  $\succ$  describes the order of moves,  $\mathcal{O}$  determines what happens at final nodes and  $u$  expresses the players preferences. The collection of information sets  $\mathcal{I}$  expresses the imperfect information by containing all decision nodes which a player cannot distinguish when making a move. Thus, in the game it is more apt to say that players move at information sets, rather than at decision nodes.

Given a game  $\Gamma$ , we can further define the notion of strategies and strategy profiles, again largely following Mas-Collel et al. (1995).



**Definition 4.** Let  $\mathcal{I}^i = \{I \in \mathcal{I} \mid \phi(I) = i\}$ , the set of all information sets to which player  $i$  is assigned. A *pure strategy*  $\sigma^i$  (or just *strategy*) for player  $i \in P$  is a mapping  $\mathcal{I}^i \rightarrow \mathcal{A}$  such that for each  $I \in \mathcal{I}^i$ ,  $\sigma^i(I) \in \alpha(I)$  (that is,  $\sigma$  must assign an action which is available at  $I$ ). A *mixed strategy* is a probability assignment on the set of pure strategies. A collection of  $\sigma_i$  for all  $i \in P$  is called a *strategy profile*.

The idea of translating a game of incomplete information into a game of imperfect information by the introduction of a nature player is due to Harsanyi (1968).

**Definition 5.** Let player  $W$  be the nature player. At each node such that  $\phi(d) = W$ , the nature player only has a single available strategy, pre-specified by the game.

In addition to a strategy profile  $\sigma$ , we also define a system which will govern the beliefs of the players throughout the game:

**Definition 6.** The *belief system*  $\mu$  specifies, at every information set  $I \in \mathcal{I}$ , a probability distribution  $\mu(I)$  over each  $d \in I$ , i.e. for all  $I \in \mathcal{I}$ ,  $\sum_{d \in I} \mu(d) = 1$  and  $\mu(d) \geq 0$  for all  $d \in \mathcal{D}$ . A *strategy profile*  $\sigma = (\sigma^1, \dots, \sigma^n)$  and a belief system  $\mu$  form an *assessment*  $(\sigma, \mu)$ .

## 2.4.2 Games on the Information Structure

The next definition describes a class of extensive form games which are compatible with the information structure described in Section 2.3.1.

**Definition 7.** Let  $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$  be the information structure as defined in Section 2.3.1 and consider a game  $\Gamma$ . The game  $\Gamma$  is a *valid game on*  $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$  if it satisfies the following description:

1. The set of players includes all individuals in  $N$  and a nature player  $W$ .
2. The game starts with a move by the nature player determining  $E$ , who plays the action  $E = e$  with probability  $\mathbb{P}(E = e)$ .

3. The collection of information sets  $\mathcal{I}$  can be partitioned into sets  $\mathcal{I}_1, \dots, \mathcal{I}_T$ , with the interpretation that for any  $I \in \mathcal{I}_t$ , the player  $\phi(I)$  moves at time  $t$ .
4. The first move of player  $W$  at  $t$  is to choose  $S_t^i = s$  for all  $i \in N$ , where for all  $s \in \mathcal{S}$ ,  $S_t^i = s$  happens with probability  $\mathbb{P}(S_t^i = s | E = e)$ .
5. At each  $I \in \mathcal{I}_t$  at which a player in  $N$  moves, a player knows all previous choices  $S_1^i, \dots, S_t^i$  by player  $W$ , but cannot distinguish different choices for  $E$  or for  $S_t^j$  for any  $j \neq \phi(I)$ .

Conditions 1 and 2 make sure that the underlying information structure of the game is given by  $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$ . Conditions (3-5) mean that the rest of the game must be compatible with the timing of the information structure and must leave the privacy of information intact. Given these restrictions, the game could still be quite different from the game played by participants of a market scoring rule. For instance, a valid game could have each player make predictions without knowing the previous probability assessments of other players. There could also be more elaborate structures in which for instance, at a time point  $t$ , individuals can “challenge” the reports made by other individuals, and then individuals make additional probabilistic reports. It is also possible that the nature player moves again, e.g. by tossing a coin to determine which player receives a payoff. Some of these other forms are discussed in Section 2.6. Further formal treatment and comparisons of market scoring rules with other valid games on  $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$  seem like a fruitful area of future research.

### 2.4.3 Equilibria and Information Revelation

A strategy profile  $\sigma$  and a belief system  $\mu$  fully determine the moves made by all players (including  $W$ ), and thus pin down a probability for any decision node to be reached starting from any information set  $I$ . Suppose that  $I \in \mathcal{I}_t$  and  $\phi(I) = i$ . Since each such

decision node corresponds to a specific signal history and event  $E = e$ , we can define a probability measure representing the beliefs of individual  $i$  at  $t$  over the signal histories and about the event  $E$ . Let  $\mu_I(\cdot|\sigma)$  denote this probability measure taken at information set  $I \in \mathcal{I}$  (conditioning on strategy  $\sigma$ ) and let  $\mu_0(\cdot|\sigma)$  denote the case that  $I \in \mathcal{I}_0$  (and thus  $I$  only contains  $d_0$ ). Using this notation, we can define the following restriction on a belief system  $\mu$ :

**Definition 8.** The assessment  $(\sigma, \mu)$  is *consistent with Bayes' rule* if for any  $I \in \mathcal{I}$  and any  $d \in I$

$$\mu_I(d|\sigma) = \frac{\mu_0(d|\sigma)}{\mu_0(I|\sigma)}.$$

In similar fashion, we can also write  $\mu_I(o|\sigma)$ , to denote the probability of outcome  $o \in \mathcal{O}$  as seen from information set  $I$ , given belief system  $\mu$  and strategy profile  $\sigma$ . We can use this notation to define our equilibrium concept:

**Definition 9.** Let  $\Gamma$  be a valid game on  $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$ . The assessment  $(\sigma, \mu)$  is a *Perfect Bayes-Nash equilibrium* (or just *equilibrium*) if  $(\sigma, \mu)$  is consistent with Bayes' rule and for any  $i \in N$ , and any  $I \in \mathcal{I}$ :

$$\int_{\mathcal{O}} u^i(o) d\mu_I(o|\sigma) > \int_{\mathcal{O}} u^i(o) d\mu_I(o|\tilde{\sigma}) \tag{2.4.2}$$

for any  $\tilde{\sigma}$  such that all of the following three conditions hold:

1.  $\tilde{\sigma}^i(I) \neq \sigma^i(I)$ ,
2.  $\tilde{\sigma}^i(I') = \sigma^i(I')$  for all  $I'$  from which  $I$  can be reached,
3.  $\tilde{\sigma}^j = \sigma^j$  for all  $j \neq i$ .

In words, at any information set  $I$ , reporting according to strategy  $\sigma^i$  leads to a higher expected payoff than a deviating strategy  $\tilde{\sigma}^i$ , as measured by  $\mu_I$  at this information set.

This is under the restriction that all other players stick to their strategy and that  $\tilde{\sigma}$  is identical to  $\sigma$  for all information sets preceding  $I$ . The latter restriction reflects the dynamic nature of the setting. It may happen that at some information set, an individual regrets having reported according to  $\sigma^i$  earlier, because new information makes  $i$  expect that different actions would have led to a higher pay-off. However, for an equilibrium strategy  $\sigma$ , *at the time of acting*, each plan of *future* actions according to  $\sigma$  has the highest expected payoff.

**Definition 10.** Consider information sets  $I, I' \in \mathcal{I}_t^i$  which only differ with respect to the signal  $S_t^i$ . An equilibrium assessment  $(\sigma, \mu)$  is *separating* if for all  $i \in N$  and all such  $I, I'$  which are reached with positive probability, we have  $\sigma^i(I) \neq \sigma^i(I')$ .

That is, if a player has the same information at information sets  $I, I' \in \mathcal{I}_t^i$ , except that according to  $I$ ,  $i$  learns at  $t$  that  $S_t^i = s$  and according to  $I'$ ,  $i$  learns at  $t$  that  $S_t^i = s' \neq s$ , then  $i$  chooses a different action at the two information sets. In a separating equilibrium assessment, the market-maker can infer the signals by observing the players' actions. Therefore, I will say in the following that:

**Definition 11.** A valid game  $\Gamma$  *induces information revelation*, if there is a separating equilibrium assessment  $(\sigma, \mu)$  in  $\Gamma$ .

#### 2.4.4 Market Scoring Rules as Games

With the notion of a valid game, we can define a *market scoring rule game* (MSR game) as a valid game in which each player moves exactly once at  $t$ , and the information set of each player at  $t$  contains all previous reports made by all other players, and the signal realizations  $s_1^i, \dots, s_t^i$ . For MSR games, we can simplify the definition of a Perfect Bayes-Nash equilibrium somewhat. Note that the probability measure  $\mu_I$  we constructed above also assigns probabilities over signal histories: for some history  $h$ , we can simply define  $\mu_I(h|\sigma)$  as the sum over the probability assigned by  $\mu_I(\cdot|\sigma)$  to all decision nodes which are

reached under history  $h$ . Note further that a strategy profile  $\sigma$  and a history  $h$  completely determine the outcome  $o \in \mathcal{O}$  in an MSR game, which I will denote  $\sigma(h)$ . Then:

*Remark 1.* An assessment in an MSR game  $(\sigma, \mu)$  is a Perfect Bayes-Nash equilibrium if it is consistent with Bayes' rule and for any  $i \in N$ , and any  $I \in \mathcal{I}^i$ :

$$\sum_{h \in \mathcal{H}} \mu_I(h|\sigma) \mathbb{E} \left[ \pi^i(\sigma(h), E) | H = h \right] > \sum_{h \in \mathcal{H}} \mu_I(h|\sigma) \mathbb{E} \left[ \pi^i(\tilde{\sigma}(h), E) | H = h \right] \quad (2.4.3)$$

for any  $\tilde{\sigma}$  which differs from  $\sigma$  at  $I$  and is identical for all information sets preceding  $I$ .

Finally, I will make the following restriction.

**Definition 12.** A *restricted* MSR game allows only probabilistic reports  $r_t^i$  such that there is  $s \in \mathcal{S}$  such that for each  $e \in \mathcal{E}$ :

$$r_t^i(e) = \frac{\mathbb{P}(S_t^i = s | E = e) r_t^{i-1}(e)}{\langle \mathbb{P}(S_t^i = s | E), r_t^{i-1} \rangle} \quad (2.4.4)$$

As explained above, this restriction simplifies the description of strategies and thereby the statement and proof of the main result.

## 2.5 The Logarithmic Market Scoring Rule Game - Results

### 2.5.1 The LMSR Game - Defining Truthfulness

Having defined a market scoring rule as a game, and simplified the notion of a Perfect Bayes-Nash equilibrium for such games, we will next analyze specific strategies in an MSR game and in particular show that “reporting truthfully” forms an equilibrium.

**Definition 13.** A report history  $r = (r_0, \dots, r_1^1, \dots, r_1^n, \dots, r_t^1, \dots, r_t^n)$  is *explained* by the signal

history  $h_t = (h_t^1, \dots, h_t^n)$  at  $(t, i)$  if

$$r_u^j = \mathbb{P}\left(E \mid H_t^j = h_u^j, H_t^{-j} = h_u^{-j}\right)$$

for all  $(u, j)$  such that (i)  $u < t$  or (ii)  $u = t$  and  $j < i$ . Furthermore, a report history  $r$  is *believable at*  $(t, i)$  if there is a signal history which explains it at  $(t, i)$ .

A restricted MSR game can thus also be defined as an MSR game in which only believable reports are allowed. Note that the set of believable reports is a finite set – and thus a set of measure zero among all possible reports. In practical implementation, this means that the market-maker needs to know the probability measure  $\mathbb{P}$  precisely. As I argue in the Discussion below, this practical limitation makes it desirable to extend the results in this chapter to market scoring rules without the restriction.

Believable report histories naturally extend to believable information sets:

**Definition 14.** Let  $r$  be the report history in all decision nodes in information set  $I \in \mathcal{I}_t$  at which player  $i$  is to move. Then  $I$  is *explained by the signal history*  $h_t$  if  $r$  is explained by  $h_t$  at  $(t, i)$ .

The next definition clarifies what it means for market participants to make truthful reports:

**Definition 15.** A strategy profile  $\sigma$  is *truthful* if for any information set  $I$  which is explained by the signal history  $h_t$

$$\sigma^i(I) = \mathbb{P}\left(E \mid H_t^{-i} = h_t^{-i}, H_{t-1}^i = h_{t-1}^i\right) \quad (2.5.1)$$

A truthful profile is thus one in which individuals always report their probabilistic beliefs, taking the signals implied by previous probabilistic reports at face value.

I also introduce the notion of a *trusting* belief profile

**Definition 16.** Let information set  $I$  be explained by the signal history  $h_t$  at  $(t, i - 1)$  and let  $h_t^i$  be  $i$ 's true signal history at  $I$ . In an MSR game, a belief profile  $\mu$  is *trusting* if for any decision node  $d \in I$  represented by  $(\tilde{h}_t^{-i}, e)$ :

$$\mu(d) = \begin{cases} \mathbb{P}(E = e | H_t^{-i} = h_t^{-i}, H_t^i = h_t^i) & \text{if } \tilde{h}_t^{-i} = h_t^{-i} \\ 0 & \text{else} \end{cases} \quad (2.5.2)$$

for any  $d \in I$ .

That is, in a trusting assessment, an individual assigns all probability mass to decision nodes consistent with the history explaining the information set. Among these, the decision nodes are assigned a probability according to the event outcome  $e \in \mathcal{E}$  that characterizes them. As there is exactly one decision node corresponding to each  $e \in \mathcal{E}$  among the decision nodes which are assigned positive probability, from

$$\sum_{e \in \mathcal{E}} \mathbb{P}(E = e | H_t^{-i} = h_t^{-i}, H_t^i = h_t^i) = 1$$

we can conclude that  $\sum_{d \in I} \mu(d) = 1$ .

In a restricted MSR game, all information sets are believable. We can then see by induction that there is only a unique assessment which is trusting and truthful: When individual 1 makes a report in period 1, the only trusting belief about the signal history is to assign probability 1 to  $h_0$ , and the only truthful report is  $\mathbb{P}(E | S_1^1 = s_1^1)$ . For the inductive step, note that if the entire signal history can be inferred before an individual  $i$  moves at  $t$ , and  $i$  makes a truthful report at  $t$ , then  $i$ 's signal at  $t$  can be inferred as well from the stochastic relevance property from Equation 2.3.3. Therefore, when the strategy is truthful, a trusting belief system must assign probability mass 1 to decision nodes representing the true signal history, according to the true conditional probability  $\mathbb{P}(E = e | H_t^{-i} = h_t^{-i}, H_t^i = h_t^i)$ . There is thus only a single trusting and truthful assess-

ment in the restricted MSR game

**Definition 17.** In a restricted MSR game, let the (unique) trusting and truthful assessment be called the *max-trust assessment*.

The max-trust assessment is consistent with Bayes' rule, since for any decision node  $d$  which can be explained by history  $h_t^{-i}$ :

$$\begin{aligned} \frac{\mu_0(d|\sigma)}{\mu_0(I|\sigma)} &= \frac{\mathbb{P}(E=e \text{ and } H_t^{-i}=h_t^{-i} \text{ and } H_t^i=h_t^i)}{\mathbb{P}(H_t^{-i}=h_t^{-i} \text{ and } H_t^i=h_t^i)} \\ &= \mathbb{P}(E=e | H_t^{-i}=h_t^{-i}, H_t^i=h_t^i) = \mu_I(d|\sigma). \end{aligned}$$

where  $\mu_0, \mu_I$  are defined as in Definition 8.

In the unrestricted MSR game without the restriction, we would also need to specify what  $\sigma$  and  $\mu$  prescribe at non-believable history nodes to complete the description of an assessment. There is then not a unique trusting and truthful assessment. For the result stated in this chapter however, Equations (2.5.1) and (2.5.2) are sufficient.

## 2.5.2 Information Revelation

It is straightforward to see that the max-trust assessment is separating. Thus, if it is a Perfect Bayes-Nash equilibrium in the restricted LMSR game, we can conclude that the restricted LMSR induces information revelation. We will next prove that this is indeed the case, the main insight of this chapter.

**Proposition 1.** *If  $\mathbb{P}$  satisfies conditional independence, the restricted LMSR induces information revelation. Specifically, the max-trust assessment is a Perfect Bayes-Nash equilibrium in the restricted LMSR game.*

*Proof.* To begin the argument, suppose that all individuals report truthfully and have trusting beliefs. Consider some  $i \in N$ . We want to show that any deviation from the truthful strategy leads to a lower expected payoff for  $i$ . The underlying idea of showing



this is as follows: The “instantaneous payoff” for  $i$  for reporting at any  $t \in \mathcal{T}$  (that is, the payoff for just this report, after the event  $E$  is observed) is given by a proper scoring rule. Since proper scoring rules score higher in expectation for truthful probabilistic reports, a profitable deviation could only come from misleading other participants, and gaining from correcting their reports at a later point in time. However, under the conditional independence information structure, the gain from correcting this belief will always be lower in expectation than the loss from the initial misreport, as we can show by making use of Lemma 1.

To make this argument precise, fix an individual  $i$ , and consider an arbitrary strategy profile  $\tilde{\sigma}$  in which  $i$  is not truthful at some information set. Fix a signal history  $\bar{h}$  which induces reports in which  $i$  is not truthful, let  $\underline{t}$  be the *first* time period in which  $i$  does not report truthfully and let  $I$  be the corresponding information set. We will construct a new strategy  $\sigma$  from  $\tilde{\sigma}$  in which  $i$  instead reports truthfully at  $t$  in history  $\bar{h}$ . Moreover,  $\sigma$  “behaves like  $\tilde{\sigma}$  everywhere else”, in the following sense. All other individuals stick to the max-trust strategy for each report which is reached with positive probability starting from  $I$  (given the strategy profile  $\tilde{\sigma}$ ). Therefore, when individual  $i$  is to move at time period  $t+1$ , the entire signal history  $H_t^{-i}$  can be inferred. Whenever two information sets  $I'$  and  $I''$  can be reached with positive probability starting from  $I$  according to  $\sigma$  and  $\tilde{\sigma}$ , respectively, we then let  $\sigma(I') = \tilde{\sigma}(I'')$  whenever the inferred signal history is identical for  $I'$  and  $I''$ . That is, we pair information sets from the path starting from  $I$  according to  $\sigma$  and information sets starting from  $I$  according to  $\tilde{\sigma}$  which have the same signal history, and set  $\sigma$ ’s report at each member of the pair equal to what  $\tilde{\sigma}$  specifies for the other member. We can leave  $\sigma$  unspecified for all other information sets, as they occur with probability zero as viewed from information set  $I$ , and thus do not affect the expected payoff as viewed from  $I$ .

To simplify notation, let  $i = 1$ . The report just before  $i$ ’s next report is thus  $r_t^n$ . As a

further simplification, let  $\underline{t} = 1$ .<sup>7</sup> Denote 1's true signal at time period 1 by  $s$ . Then, we have, for some,  $\tilde{s} \in \mathcal{S}$ ,  $\tilde{s} \neq s$ :

$$\sigma(I_1^1) \equiv r_1^1 = \mathbb{P}(E | S_1^1 = s) \neq \tilde{\sigma}(I_1^1) = \mathbb{P}(E = e | S_1^1 = \tilde{s}) \equiv \tilde{r}_1^1 \quad (2.5.3)$$

Note next that by construction the payoff difference for strategy  $\tilde{\sigma}^i$  and strategy  $\sigma^i$  is only determined by the score  $i$  receives for the report in period 1 and the score  $i$  has to pay in period  $t + 1$ .

In expectation, the payoff difference for strategy  $\tilde{\sigma}^i$  and strategy  $\sigma^i$  is given by:

$$\sum_{h \in \mathcal{H}} \mu_I(h | \sigma) \mathbb{E}[\pi^1(\sigma(h), E) | H = h] - \sum_{h \in \mathcal{H}} \mu_I(h | \sigma) \mathbb{E}[\pi^1(\tilde{\sigma}(h), E) | H = h] \quad (2.5.4)$$

We further have that the payoffs from  $\sigma$  and from  $\tilde{\sigma}$ , as viewed from information set  $I$ , only differ at periods 1 and 2. Under strategy  $\sigma$ , if the event realization is  $e \in \mathcal{E}$ , individual 1 receives a payoff of  $\log(r_1^1(e)) - \log(r_0(e))$  for the report in period 1 and a payoff of  $\log(r_2^1(e)) - \log(r_1^n(e))$  for the report in period 2. Writing reports made under strategy  $\tilde{\sigma}$  as  $\tilde{r}_t^j$ , individual 1 receives  $\log(\tilde{r}_1^1(e)) - \log(r_0(e))$  and  $\log(\tilde{r}_2^1(e)) - \log(\tilde{r}_1^n(e))$  under strategy  $\tilde{\sigma}$ . Note that by construction of the strategy  $\sigma$ , the reports  $\tilde{r}_2^1(e)$  and  $r_2^1(e)$  are identical. The expression (2.5.3) thus reduces to:

$$\sum_{e \in \mathcal{E}} \mathbb{P}(E = e | S_1^1 = s) \sum_{h \in \mathcal{H}_1^{-1}} \mathbb{P}(H_1^{-1} = h | E = e, S_1^1 = s) \left[ \log\left(\frac{r_1^1}{\tilde{r}_1^1}\right) - \log\left(\frac{r_1^n}{\tilde{r}_1^n}\right) \right]$$

---

<sup>7</sup>Since all reports are truthful up until  $\underline{t}$ , we could instead define  $\tilde{\mathbb{P}} = \mathbb{P}(\cdot | H_{\underline{t}} = h_{\underline{t}})$ . Since  $\tilde{\mathbb{P}}$  is a probability measure and satisfies conditional independence, we could replace  $\mathbb{P}$  by  $\tilde{\mathbb{P}}$  in all steps of the proof. Moreover, instead of taking expectation over all histories, we would take expectation over all histories  $h$  which agree with  $\bar{h}$  up until  $\underline{t}$ . Setting  $\underline{t} = 1$ , avoids this notational complexity while leaving all arguments intact.

$$= \sum_{e \in \mathcal{E}} \mathbb{P}(E = e | S_1^1 = s) \sum_{h \in \mathcal{H}_1^{-1}} \mathbb{P}(H_1^{-1} = h | E = e) \left[ \log \left( \frac{r_1^1(e)}{\tilde{r}_1^1(e)} \right) - \log \left( \frac{r_1^n(e)}{\tilde{r}_1^n(e)} \right) \right] \quad (2.5.5)$$

where

$$r_1^n(e) = \mathbb{P}(E = e | H_1^{-1} = h, S_1^1 = s),$$

and

$$\tilde{r}_1^n(e) = \mathbb{P}(E = e | H_1^{-1} = h, S_1^1 = \tilde{s}).$$

Invoking Lemma 1, we get that

$$\log \left( \frac{r_1^1(e)}{\tilde{r}_1^1(e)} \right) - \log \left( \frac{r_1^n(e)}{\tilde{r}_1^n(e)} \right) = \log \left( \frac{\mathbb{P}(H_1^{-1} = h | S_1^1 = s)}{\mathbb{P}(H_1^{-1} = h | S_1^1 = \tilde{s})} \right)$$

and hence the expression in (2.5.5) equals

$$\begin{aligned} & \sum_{e \in \mathcal{E}} \mathbb{P}(E = e | S_1^1 = s) \sum_{h \in \mathcal{H}_1^{-1}} \mathbb{P}(H_1^{-1} = h | E = e) \log \left( \frac{\mathbb{P}(H_1^{-1} = h | S_1^1 = s)}{\mathbb{P}(H_1^{-1} = h | S_1^1 = \tilde{s})} \right) \\ &= \sum_{h \in \mathcal{H}_1^{-1}} \mathbb{P}(H_1^{-1} = h | S_1^1 = s) \log \left( \frac{\mathbb{P}(H_1^{-1} = h | S_1^1 = s)}{\mathbb{P}(H_1^{-1} = h | S_1^1 = \tilde{s})} \right) \\ &= L \left( \mathbb{P}(H_1^{-1} = h | S_1^1 = s), \mathbb{P}(H_1^{-1} = h | S_1^1 = \tilde{s}) \right) > 0. \end{aligned}$$

We have thus shown that the expected payoff for following strategy  $\sigma$  is higher than for following strategy  $\tilde{\sigma}$ . Recall that we considered an arbitrary history  $\bar{h}$  in which  $\tilde{\sigma}$  specified at least one untruthful probability report. In this history  $\bar{h}$ , there can only be a finite number of misstated probabilistic beliefs. Let  $l$  be this number. We can construct a sequence  $\sigma^1, \dots, \sigma^l$  where  $\sigma^1 = \sigma$  and for  $k > 1$ ,  $\sigma^k$  replaces the first lie of strategy  $\sigma^{k-1}$  by a truthful report. Since each such  $\sigma^k$  is a deviation from the truthful strategy in  $\bar{h}$ , it follows from the above that  $\sigma^k$  has a higher expected payoff than  $\sigma^{k-1}$ . Consequently,  $\sigma^l$ , which is truthful throughout  $\bar{h}$ , has the highest expected payoff. Since we are showing

this relation for an arbitrary history  $\bar{h}$ , we can infer that the truthful strategy improves upon any deviation  $\tilde{\sigma}$ , at all signal histories. This proves the proposition.  $\square$

## 2.6 Discussion

The main result of this chapter is that market scoring rules are robust to strategic manipulation, even if individuals update on the information provided by other market participants and by new information they receive privately – and expect other participants do to so as well. In previous work, individuals are either considered 'myopic', i.e. they disregard information provided by other participants, or individuals receive information only once, at the beginning of the market.

As a further, more technical, contribution of this chapter, I introduced a distinction between two versions of market scoring rules. In the *restricted* market scoring rule, the market-designer knows the underlying information structure of the setting, and can rule out probabilistic reports which cannot possibly reflect market participants' subjective beliefs (because there is no information they could have received which would make it their subjective belief). Participants can still mislead the market-maker or other individuals about their private information, but in doing so, they need to pretend to have received some information which occurs with positive probability. In an *unrestricted market* such reports are allowed, and the designer does not need to know the information structure. Restricting the market scoring rule simplifies the description of equilibrium strategies and allows to highlight the main mechanics of how the LMSR induces truth-telling. However, in practice, the market-designer may not know the distribution of private information. A market which only allows specific types of reports could also be seen as intransparent. Therefore, it seems preferable to use the LMSR in an unrestricted way, and indeed it has been in all practical applications. A particularly important avenue of theoretical investigation would therefore be to study the properties of the unrestricted LMSR.

A crucial assumption in deriving the main result is that individuals' signals – their privately held information – is independent conditional on the outcome of the true event. To further illustrate what is entailed by this assumption, I will next give an example in which it is violated. Suppose we have  $\mathcal{E} = \{1, 2\}$ ,  $\mathcal{S} = \{1, 2\}$ ,  $T = 2$  and  $\mathbb{P}(E = 1) = \frac{1}{2}$ . Suppose now that

$$\mathbb{P}(S_1^1 = 0) = 1$$

That is, with certainty individual 1 does not receive an informative signal in period 1.<sup>8</sup> Moreover, we have:

$$\begin{aligned} \mathbb{P}(S_1^2 = 1, S_2^1 = 1 | E = 1) &= \frac{1}{2} - \varepsilon, & \mathbb{P}(S_1^2 = 1, S_2^1 = 2 | E = 1) &= \frac{\varepsilon}{2}, \\ \mathbb{P}(S_1^2 = 2, S_2^1 = 2 | E = 1) &= \frac{1}{2}, & \mathbb{P}(S_1^2 = 2, S_2^1 = 1 | E = 2) &= \frac{\varepsilon}{2}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}(S_1^2 = 1, S_2^1 = 2 | E = 0) &= \frac{1}{2} - \frac{\varepsilon}{2}, & \mathbb{P}(S_1^2 = 1, S_2^1 = 1 | E = 0) &= \frac{\varepsilon}{2}, \\ \mathbb{P}(S_1^2 = 2, S_2^1 = 1 | E = 0) &= \frac{1}{2} - \frac{\varepsilon}{2}, & \mathbb{P}(S_1^2 = 2, S_2^1 = 2 | E = 0) &= \frac{\varepsilon}{2}. \end{aligned}$$

where  $\frac{1}{2} > \varepsilon > 0$ . Note that for small  $\varepsilon$ , the state  $E = 1$  thus corresponds to signal distributions in which the signals  $S_1^2$  and  $S_2^1$  are almost always identical, and in the state  $E = 0$  the signals are almost always different from each other. Realizations of the two signals are then highly informative of each other, *especially* when conditioning on the true event. We have

$$\mathbb{P}(E = 1 | S_1^2 = 1) = \frac{1}{2} - \frac{\varepsilon}{2} = \frac{1}{2}(1 - \varepsilon)$$

$$\mathbb{P}(E = 1 | S_1^2 = 2) = \frac{\frac{1}{2}}{1 + \frac{\varepsilon}{2}} = \frac{1}{2 + \varepsilon}$$

For small  $\varepsilon$ , the information contained in the individual signal  $S_1^2$  does thus not move the posterior of individual 2 far away from the prior of  $\frac{1}{2}$ . However, if individual 2 were to know  $S_1^2$  and  $S_2^1$ , the situation is quite different: if  $S_2^1 = S_1^2$  then  $E = 1$  with near

---

<sup>8</sup>With this assumption, the game essentially becomes the “Sequence Selection Game” considered in Chen et al. (2010).

certainty and else  $E = 0$  with near certainty. This configuration means that the max-trust assessment is not an equilibrium. We assumed that individual 1 does not learn anything in period 1. Thus, in period 1, individual 1 reports  $r_1^1 = 0.5$ . Next, individual 2 is to move. Suppose individual 2 observes  $S_1^2 = 1$ . Truth-telling would require reporting  $r_1^2 = \frac{1}{2}(1 - \varepsilon)$ . If instead 2 reports  $\frac{1}{2+\varepsilon}$  however, individual 1 would conclude that individual 2 observed  $S_1^2 = 2$ . Then, if  $S_2^1 = 1$ , individual 1 would conclude that  $E = 1$  and if  $S_2^1 = 2$ , then individual 1 would conclude that  $E = 2$  – both with near certainty – and report accordingly. In both cases, individual 2 can infer  $S_2^1$  and would conclude that the state is in fact the opposite. Individual 2 can then make a large expected profit by reporting in period 2. In contrast, the expected loss from misreporting in period 1 is only  $L\left(\frac{1}{2}(1 - \varepsilon), \frac{1}{2+\varepsilon}\right)$  – which can be arbitrarily close to 0 for sufficiently small  $\varepsilon$ .

Note that this example is quite drastic. Where conditional independence stipulates that upon knowing the event  $E$ , knowing  $S_1^2$  adds no information about the signal  $S_2^1$  (and thus knowing the event  $E$  annihilates the predictive value of  $S_1^2$  for  $S_2^1$ ), in the example knowing the event  $E$  *increases* the ability to predict  $S_2^1$  from  $S_1^2$ . It is difficult to imagine practical settings in which this is a plausible assumption. It can further be shown that Proposition 1 continues to hold if one assumes that knowing the event  $E$  never increases the ability to predict  $S_2^1$  from  $S_1^2$ . However, if conditional independence is replaced by this more general assumption, the updating formula given by Equation (2.3.4) does not hold anymore. The current market price does then not reflect all relevant information – participants need to know the entire report history to compute their posterior beliefs. Thus, the fundamental property of prediction markets to have prices reflect all relevant information is quite closely linked to the assumption of conditional independence, and should thus only be applied when conditional independence can be plausibly assumed.

Another notion which is discussed in studies of the myopic strategic properties of MSRs, but which has been completely neglected in this chapter so far, is “sybil-proofness” (Douceur, 2002). A sybil is an additional participant in the market, who is in fact the

same person as an already existing member. Sybil-proofness is important in an online setting in which one cannot verify the identity of every new member, and there is value in having a large number of participants. Both of these requirements seem reasonable in the context of prediction markets. Suppose that the set of individuals  $N$ , and their labels  $1, \dots, n$  are fixed. If a new entrant gets the label  $n + 1$ , then it is easy to see that the LMSR is sybil-proof: When the sybil is to move, it is best – as per Proposition 1 – to report the true belief, which will simply be the previous report, as the sybil does not in fact have any new information. The payoff of the sybil is then zero, and there is no effect on other market participants. A different situation arises when the order is not fixed, but randomly determined at each time period, or determined after individuals had the possibility to create sybils. In this case, by creating an additional market participant, an individual  $i$  can increase the chance to make a report earlier within the same period. When the other market participant controlled by  $i$  is to move,  $i$  can simply let the report equal the preceding report – ensuring a payoff of zero for this report. For the same reason that the LMSR is robust to strategic misstatements, the LMSR with a randomly determined order is susceptible to sybil-attacks: it is attractive for market participants to report as early as possible.

The problem of sybil-attacks is not severe if signals occur relatively rarely over time. In this case, the market-maker can choose  $\mathcal{T}$  to be very fine-grained (allowing reports at every millisecond, say). By default, market participants would be “inactive”, i.e. just reporting what the previous report said. The market-maker could then charge a minimal fee for entering the market. For a very fine-grained  $\mathcal{T}$ , the benefit of reporting early within a time-period decreases, since for most periods, most participants will not make “active” reports (changing the probability of the previous report). In expectation, creating a sybil may then be too costly to be profitable.

As a result, the LMSR can deal with the problem of sybil attacks much more efficiently than a rival incentive scheme in which each individual simply is rewarded by a proper scor-

ing rule for their individual reports. As was noted in the introduction, such a scheme may lead to unlimited costs for the market-maker, as the number of participants increases. This could be counteracted by dividing every participants' payoff by the number of participants. However, such a scheme would still be open to sybil-attacks, as participants could essentially double their payments by creating a sybil (and triple by creating two sybils, and so on). Moreover, multiplying payments by  $\frac{1}{n}$  has the disadvantage that even very informative reports might lead to a very small payoff. This way, unlike in the LMSR, the payoff of a very well-informed individual could be drastically reduced by a large crowd of noisy market participants. It seems like a fertile area of future research to characterize other valid payment schemes on the information structure which have similar properties with respect to sybil-attacks and payments for information, and which properties fully characterize the LMSR.

Finally, it would be of interest to loosen the assumption of risk neutrality. Risk neutrality is assumed in most of the prediction market literature, and can sometimes be justified through the observation that the behavior of large financial players can be approximately described by risk-neutrality. However, in case of the LMSR it is an explicit goal to aggregate information from a large number of participants, some of which may be quite budget-constrained. Therefore, extensions of the LMSR which make it robust to risk aversion seem desirable. This would open up connections to a large literature on belief elicitation for decision-makers who are not risk-neutral, and may not behave according to expected utility theory (e.g. Kothiyal et al., 2010; Lambert, 2018; Offerman et al., 2009). In principle, it should be possible to replace the scoring rules in the LMSR by these more general individual schemes. However, if one keeps the goal of a fixed loss to the market-maker, and the assumption that informative reports should not decrease in value simply because the number of market participants increases, bad forecasts need to be inevitably connected to a financial loss some of the time. This puts some limits on what can be achieved for risk averse decision-makers. What can and cannot be achieved



by LMSRs which are adapted for non-risk neutral decision-makers,<sup>9</sup> is yet to be spelled out.

## 2.7 Concluding Remarks

The model provided in this chapter gives a setting in which the LMSR induces information revelation. I have further outlined some fruitful avenues of future research which could make LMSRs applicable to a larger class of decision-makers – and to a wider class of practical applications. Given the promising aspects of LMSR, it is encouraging that it is already employed in practice (with Dreber et al., 2015 and Camerer et al., 2016, 2018 as particularly noteworthy applications). This gives rise to the hope that in the near future, prediction markets to aggregate beliefs about verifiable, near term, events may be in widespread use, and that these markets will inform and improve public policy-making.

---

<sup>9</sup>As an example how this could work, consider the classic idea to transform payoffs from a scoring rule into probabilities of winning a fixed prize. Due to the bounded loss assumption, it seems necessary that respondents not only win the prize some of the time, but also need to pay it to other participants. For some decision-makers, incentive compatibility will be restored if they are allowed to determine the size of the payoffs (essentially picking a scalar  $a \in [0, 1]$  with which to multiply the prize). Future work could precisely state the class of decision-makers for which such adjustments are possible.



# Chapter 3

**Follow the money, not the majority:**

**Incentivizing and aggregating**

**expert opinions with Bayesian markets<sup>1</sup>**

## 3.1 Introduction

A centerpiece of economic theory is the idea that markets are efficient aggregators of information (Hayek, 1945; Hurwicz, 1960; Fama, 1970). Historically, this idea has been a descriptive one, explaining the success of market institutions as we encounter them in the real world. In line with an ongoing shift from mere description towards applying economic theory to also create institutions (Roth, 2002, 2018), economists have more recently argued for the use of artificially designed markets (“prediction markets”) with information aggregation as a designated goal (Arrow et al., 2008; Hanson, 2003, 2013). Successful applications range from forecasts of political elections (Forsythe et al., 1992; Berg et al., 2008) to business sales (Cowgill and Zitzewitz, 2015; Gillen et al., 2017) and the replicability of experiments in social science (Dreber et al., 2015; Camerer et al.,

---

<sup>1</sup>This chapter is based on an eponymous paper co-authored with Aurélien Baillon and Tong V. Wang.

2016, 2018). For those tools to be successful it is however necessary, both for theoretical and practical reasons, that the true answer to the question they are applied to can be determined within a relatively short time frame. This poses a challenge when we wish to apply them to questions such as whether extraterrestrial life exists or whether genetically modified food is dangerous. For such questions, not only is the answer presently unknown, but it is also uncertain when and how the answer will be known, if at all. This creates what we call the “incentive problem”: When asking experts (or a crowd of laypeople) to give an answer, how can we incentivize their truth-telling if we do not know whether or when the correct answer will be known?

Besides the incentive problem, we furthermore face an “aggregation problem”: which opinion to select if experts disagree? The obvious candidate is the majority opinion, but there is no guarantee that it is the best approach. For instance, imagine each expert can design and run an experiment to test whether a statement is true. Running an experiment can be seen as drawing a binary signal (“support” or “falsify”) about the state of nature (whether the statement is true or not). In some extreme cases, a single falsification among many attempts should lead to the rejection of the statement. This would be the case under a strictly Popperian scientific methodology (Popper, 1959) or when validating a mathematical statement, where a single counterexample would be sufficient to establish its falsity. Obviously, in most scientific endeavours, especially in the social sciences, experiments can be noisy and one may expect some experiments falsifying a statement even if it is true. However, the main argument remains: some opinions, based on signals that are more difficult to get, should drive the conclusion even if they are a minority.

In this chapter, we study a mechanism which solves both the incentive problem and the aggregation problem simultaneously. We design a market in which experts report their opinions about a statement (endorse it or not) and trade an asset whose value is determined by the total endorsements. Those who endorse the statement are offered to buy the asset

from a *center* at price  $p$ , where  $p$  is randomly drawn. Essentially, buying the asset is betting that more than  $p\%$  of others will endorse the statement. Those not endorsing the statement can sell the asset to the center. Baillon (2017) showed that such a “Bayesian market” provides incentives to report opinions truthfully, avoiding the no-trade theorem (Milgrom and Stokey, 1982) through the intermediary role of the center. By making a small adjustment to this mechanism—making the price individualized, independently drawn for each market participant—we show that Bayesian markets have desirable properties with respect to aggregation as well. With sufficiently many participants, experts with the signal that indicates the actual state of nature, and only them, will make a profit. Hence, by “following the money”, we can infer the state of nature without relying on what the majority thinks.

The intuition of our result is based on an argument put forward by Prelec et al. (2017). If signals are correlated with the states of nature, there will be more signals supporting a state of nature when this state is the actual one than when it is not, and therefore, than we would have expected ex ante. Prelec et al. (2017) proposed the *surprisingly popular algorithm* (SPA) in which people are asked to endorse a state and predict the rate of endorsement. The algorithm picks the state that is more often endorsed than people predicted. Prelec et al. (2017) demonstrated theoretically and experimentally that this approach improves upon majority and confidence-weighted aggregation.

Bayesian markets allow us to obtain the same improvement but with less information and requiring less cognitive efforts from the participants. We can estimate people’s predictions by fitting supply and demand curves for the asset. Furthermore, our method simultaneously provides incentives to truthfully report opinions even if the state of nature is unobservable. Our market approach is not only simpler for participants than the method of Prelec et al. (2017). It also opens up the possibility of continuous markets, extending prediction markets to unverifiable events.

The next section of the chapter introduces the theoretical setting and the market. We

analyze payoffs at the equilibrium and show how the endorsement of those with positive payoffs indicates the actual state of nature. If the statement is true, those endorsing it can make a profit from betting on others' endorsement rate. If it is not true, those rejecting it can make a profit. The profits realize even in the absence of verification of the actual state of nature, because bets are based on endorsement rates, not on states.

Section 3.3 describes an experiment we ran on a large sample of US students. We used a task developed for the experiment reported in Chapter 4 that ensures that the informational assumptions of the model are satisfied. Under these assumptions, homo economicus would behave exactly as our model predicts. Our experiment allowed us to test whether our method also worked for homo sapiens, without having to worry whether the informational part of the model perfectly described the reality. We compared our method to the majority opinion and to the SPA. Despite using less information than the SPA, our method had comparable accuracy rates. Both methods substantially improved upon the majority opinion.

## 3.2 Theory

### 3.2.1 Setting

Let  $\{Y, N\}$  be the *state space*, with  $Y$  and  $N$  the two possible *states of nature*. For instance, these two states can represent whether a statement is true or not. Which state  $S$  we are in, is assumed to be unobservable.

A group of  $n \geq 4$  expert *agents* however has private information about the state.<sup>2</sup> The *common prior* of the agents is that the probability of state  $Y$  is  $r$ . Each agent gets a *private signal*  $s_i \in \{0, 1\}$ , with the following sampling probabilities:  $P(s_i = 1 | Y) = \omega_Y$ ,  $P(s_i = 1 | N) = \omega_N$ , and  $\omega_Y > \omega_N$ . This implies that signals are informative about the

---

<sup>2</sup> $n \geq 4$  is required for technical reasons (Baillon, 2017).

state of nature,  $s_i = 1$  providing support for  $Y$  and  $s_i = 0$  for  $N$ . We do not require  $\omega_Y > 0.5 > \omega_N$ , which would be necessary for the majority of signals to be correct (in an infinite group of agents). The assumption  $\omega_Y > \omega_N$  is as mild as can be. Equality would mean that  $s_i$  is non informative and therefore, all agents would stick to the prior belief  $r$ . The opposite inequality would simply change the interpretation of the signal ( $s_i = 0$  providing support for  $Y$  and  $s_i = 1$  for  $N$ ). Together, we call the triplet  $\langle \omega_Y, \omega_N, r \rangle$  a *signal technology*.

Using Bayesian updating, agents form posterior beliefs about the actual state according to

$$r_1 \equiv P(Y | s_i = 1) = \frac{r\omega_Y}{r\omega_Y + (1-r)\omega_N}; \quad (3.2.1)$$

$$r_0 \equiv P(Y | s_i = 0) = \frac{r(1-\omega_Y)}{r(1-\omega_Y) + (1-r)(1-\omega_N)}. \quad (3.2.2)$$

For simplicity, we assume that  $\omega_Y$ ,  $\omega_N$ , and  $r$  are such that  $r_1 > 0.5$  and  $r_0 < 0.5$ . It allows us to equate an agent's signal with the state the agent believes more likely to be the actual state. If this assumption is not satisfied, signals would be informative but a single signal would not suffice to reverse one's belief. A sufficient condition for this assumption is  $r = 0.5$ , as used in our experiment.

Apart from the agents' posterior beliefs about states, we can also infer posterior expectations about the proportion of agents who received signal 1 in the population. We denote the actual value of this proportion by  $\omega$ . Since the expectation of a proportion under random sampling equals the sampling probabilities, agents who received signal 1 expect  $\omega$  to be

$$\bar{\omega}_1 \equiv E[\omega | s_i = 1] = r_1 \omega_Y + (1 - r_1) \omega_N, \quad (3.2.3)$$

whereas agents with signal 0 expect

$$\bar{\omega}_0 \equiv E[\omega \mid s_i = 0] = r_0 \omega_Y + (1 - r_0) \omega_N. \quad (3.2.4)$$

A *center* wants to find out which state we are in (the *actual* state). This center can be a policy maker consulting experts, but could just as well be an employer querying employees or a scientific association surveying its members. We make the usual assumption that the signal technology is common knowledge among the agents. However, as in Prelec (2004), Baillon (2017) and Prelec et al. (2017), the center does not know the signal technology.

Expressed in the terms of our model, the incentive and aggregation problem can be stated as follows. Each agent will report an endorsement  $e_i$ , where  $e_i = 1$  denotes that agent  $i$  endorses state  $Y$  and  $e_i = 0$  that  $i$  endorses state  $N$ . The center wants to reward the agents in such a way that it becomes profitable for them to endorse a state if and only if they believe it more likely to be the actual one. Furthermore, upon learning the endorsements  $e_1, \dots, e_n$ , the center selects one of the two states, and wishes to maximize the probability that it is the actual one. Since the state  $S$  is unobservable and the signal technology is unknown to the center, it is not possible to make the payments or selection of a state dependent on the actual state, nor the selection of the state dependent on the parameters  $\omega_Y$  and  $\omega_N$ . Thus, it is impossible to use traditional methods to elicit agents' signals or beliefs because the signals are private (impossible to directly reward truth-telling) and the beliefs are about unverifiable states  $Y$  and  $N$  (bets and scoring rules cannot be applied). Second, even knowing signals or beliefs would not enable the center to determine the state of nature because the center does not know the values for  $\omega_Y$  and  $\omega_N$ . In other words, for anyone unaware of the signal technology, observing 20% of signal 1 does not say which state we are in.

The next subsection introduces the mechanism of our solution concept, called the Bayesian market. Subsection 3.2.3 presents the underlying idea for an infinitely sized



group of expert agents, and addresses the incentive problem and the aggregation problem. It allows us to simplify many practical aspects of the mechanism; for instance, sample proportions and probabilities are equated, with  $\omega = \omega_Y$  in state  $Y$  and  $\omega = \omega_N$  in state  $N$ . In Subsection 3.2.4 we restate the results for a finite group of agents. While many of our results translate to the finite case, our approach still works better the larger the group of experts is.

### 3.2.2 Bayesian market

The center and each agent  $i$  trade an asset whose *settlement value*  $v_i$  is defined as the share of agents other than  $i$  endorsing state  $Y$ , i.e.,

$$v_i = \frac{\sum_{j \neq i} e_j}{n-1},$$

where excluding agent  $i$ 's own report prevents agent  $i$  from influencing the asset value. The center organizes a *Bayesian market* for these assets:

1. Agents simultaneously report  $e_i$  to the center only.
2. For each agent  $i$ , the center draws<sup>3</sup> a price  $p_i$  from a uniform distribution over  $(0, 1)$  and proposes the following trade to the agent, and the agent can decide to take up the offer ( $d_i = 1$ ) or not ( $d_i = 0$ ):
  - (a) If  $e_i = 1$ , agent  $i$  can buy the asset at price  $p_i$  from the center;
  - (b) If  $e_i = 0$ , agent  $i$  can sell the asset at price  $p_i$  to the center.
3. All endorsements  $e_i$  and buying/selling decisions  $d_i$  are revealed.
  - (a) If an agent decides to buy at price  $p_i$ , then there is trade under two conditions:

---

<sup>3</sup>It is important that the agents are convinced that the prices are independent of their report. To ensure it, the center may draw the prices before step 1, seal them in an envelope, and open it in step 2.

- (i) There exists another agent  $j$  selling at  $p_j \leq p_i$  and (ii) there is at least one other agent  $k$  who also endorses state  $Y$ .
- (b) If an agent decides to sell at price  $p_i$ , then there is trade under two conditions:
  - (i) There exists another agent  $j$  buying at  $p_j \geq p_i$  and (ii) there is at least one other agent  $k$  who also endorses state  $N$ .
- 4. Those agents who bought the asset collect  $v_i$  and pay  $p_i$ ; those who sold it collect  $p_i$  and pay  $v_i$ .

Step 2 differs slightly from the mechanism proposed in Baillon (2017) in which a single price  $p$  is drawn for all agents. The motivation for the change is to learn as much as possible from the decisions of different agents. When only a single price is drawn and, e.g., all potential buyers reject the trade, the center only learns that the price was larger than the buyers' reservation price, but not by how much. An alternative would be to directly ask agents for their reservation prices. The center could then draw only one random price  $p$  for all agents. This would correspond to the Becker-DeGroot-Marshak mechanism (Becker et al., 1964), but with the trading rule (step 4) in place. The advantage of binary decisions in step 2 is that they require less information from the agents, and therefore less cognitive effort. It is easier to buy/sell at a given price (equivalently, to take/reject a bet on the asset value) than to report a reservation price.<sup>4</sup>

Our mechanism as stated induces a game played among the agents. In this game, a *strategy profile* is a collection  $(e, d) = ((e_1, d_1), \dots, (e_n, d_n))$ , where  $e_i$  determines which state individual  $i$  is going to endorse depending on the signal  $s_i$ , and the trading strategy  $d_i$  assigns to each possible signal a range of prices in the  $(0, 1)$ -interval which  $i$  is going to accept when receiving a buy or sell offer from the center. Note that this definition of

---

<sup>4</sup>Asking for reservation prices, however, has advantages regarding the logistical aspects of practical implementation: In our design, a random price must be drawn for every respondent. When asking for reservation prices, respondents can be contacted by a "pen and paper" survey in which they submit their reservation prices and a public price is later credibly drawn. Whether these practical considerations outweigh the cognitive simplicity of a binary decision, will depend on the application.

strategies precludes mixed strategies and the existence of an external coordination device among agents, so that the actual endorsements made by agents are fully determined by their signal and strategy. In Section 3.5, we discuss this strategy restriction in light of our empirical results.

The mechanism assigns a payoff  $U_i(e, d)$  to each agent. Importantly, these payoffs cannot depend on the actual state of nature  $S$  or its  $\omega_S$ . A *Bayesian Nash equilibrium* of the induced game means that, conditioning on their signal, no agent expects a higher payoff by moving to another strategy, i.e.,

$$E[U_i(e, d) | s_i] \geq E[U_i((e_1, d_1), \dots, (e'_i, d'_i), \dots, (e_n, d_n)) | s_i]$$

for any  $(e'_i, d'_i) \neq (e_i, d_i)$  and all signal realizations  $s_i \in \{0, 1\}$ . We further say that a strategy profile is *truthful* if  $e_i(1) = 1$  and  $e_i(0) = 0$  for any agent  $i$ .

We assume that all agents are risk-neutral and care only about their own monetary payoff, so that  $U_i(e, d)$  is just  $i$ 's monetary payoff. If  $e_i = 1$ , agent  $i$  is potentially a buyer, and we denote by  $\pi_1(v_i, p_i)$  agent  $i$ 's monetary payoff if deciding to buy ( $d_i = 1$ ), as a function of the asset value  $v_i$  and individualized buying price  $p_i$ . Then

$$\pi_1(v_i, p_i) = \begin{cases} v_i - p_i & \text{if trade happens;} \\ 0 & \text{otherwise.} \end{cases} \quad (3.2.5)$$

Symmetrically,  $\pi_0(v_i, p_i)$  denotes agent  $i$ 's monetary payoff as a potential seller if deciding to sell ( $d_i = 0$ ):

$$\pi_0(v_i, p_i) = \begin{cases} p_i - v_i & \text{if trade happens;} \\ 0 & \text{otherwise.} \end{cases} \quad (3.2.6)$$

We will first present the intuition of Bayesian markets through the case of an infinite population.

### 3.2.3 The case of an infinite population

Three simplifications come with an infinite group of expert agents, which together, imply that the asset value is simply  $\omega_Y$  or  $\omega_N$  at the truth-telling equilibrium. First, with  $n$  infinite, the proportion of a signal in the population naturally equates the probability to get that signal. Second, excluding the agent's own signal or not from the asset value has no impact, and therefore, the asset individualization becomes equivalent to all agent trading the same asset. The third simplification is related to the trading conditions in step 4 of the Bayesian market definition. That someone else is accepting to buy or to sell at the same price is still important but the information that someone *could* make such a choice becomes trivial in an infinite group. There will always be at least one other experts receiving the same signal and one with the opposite signal. Moreover, for any nondegenerate proportion of agents endorsing each signal, there will also always be someone being offered any possible price. Hence, trade happening does not bring more information about the signal distribution than  $s_i$  does, unlike in a finite group of agents. Fortunately, we will see in the next subsection that what there is to learn is negligible when  $n$  is finite but sufficiently large.

With the three simplifications in mind, we first address the incentive problem by the following proposition.

**Proposition 2.** *Let  $\langle \omega_Y, \omega_N, r \rangle$  be a signal technology and  $n$  infinite. In the game induced by the Bayesian market, truth-telling is a Bayesian Nash equilibrium in which agents' betting strategies are such that:*

- (i) *agents whose signal is 1 buy the asset if and only if  $p_i \leq \bar{\omega}_1$ ;*
- (ii) *agents whose signal is 0 sell the asset if and only if  $p_i \geq \bar{\omega}_0$ .*

*Proof.* The main result in Baillon (2017) is essentially unaffected by the introduction of individualized prices. To get an intuition for the result, we can inspect Equations (3.2.3) and (3.2.4). It is immediate that  $\bar{\omega}_0 < \bar{\omega}_1$  since  $r_0 < r_1$  and  $\omega_N < \omega_Y$ . Thus, signal-1 agents

expect more signal-1 agents than signal-0 agents do. Consider then agent  $i$  with  $s_i = 1$  and assume all other agents are telling the truth, such that the asset value  $v_i$  equals the true share of signal-1 agents in the population. Agent  $i$  expects  $v_i$  to be  $\bar{\omega}_1$ . For  $p_i$  less than  $\bar{\omega}_1$ , agent  $i$  will be willing to buy the asset. Agent  $i$  also knows that no one would buy it at a higher price (so  $i$  has no reason to pretend to be a seller) but that some agents will be willing to sell at prices between  $\bar{\omega}_0$  and  $\bar{\omega}_1$ . For this price range, agent  $i$  foresees a profit and has the incentives to endorse  $e_i = 1$  to become a buyer. Outside this range, no trade will go through. The case  $s_i = 0$  is symmetric.  $\square$

The fact that agents trade an asset whose value they disagree on may raise the question why the no-trade theorem (Milgrom and Stokey, 1982) is not applicable here. The reason lies in the role of the center: For a trade to go through, it is sufficient that there exists a single agent who was willing to take the opposite bet. The center will verify this condition for each individual bettor, without providing further information about who the agent with the opposite bet is. Since  $0 < \omega_N < \omega_Y < 1$ , agents already know that there must be at least one disagreeing agent and thus the occurrence of trade does not provide further information about the actual  $\omega$ . Since trades are facilitated by the center,<sup>5</sup> the agents remain uncertain about the share of other agents disagreeing with them, which makes our setting different to the settings in Aumann (1976) or Milgrom and Stokey (1982) in which disagreement is impossible.

In the following proposition, we consider the aggregation problem and derive what conclusions the center can draw in the truth-telling equilibrium.

**Proposition 3.** *If  $n$  is infinite and the Bayesian market is at the truth-telling equilibrium, at least one agent has a positive payoff, all those with positive payoffs have endorsed the actual state, and all those with negative payoffs have endorsed the opposite state.*

*Proof.* At the truth-telling equilibrium, the settlement value  $v_i$  is  $\omega_N$  in state  $N$  and  $\omega_Y$  in

---

<sup>5</sup>The center will typically incur a loss from this role. The mechanism is thus not budget-balanced.

state  $Y$ . And according to Proposition 2, trades only occur for prices in the range  $[\bar{\omega}_0, \bar{\omega}_1]$ . Hence agents' payoffs, defined in Equations (3.2.5)-(3.2.6), can be simplified as

$$\pi_1(v_i, p_i) = -\pi_0(v_i, p_i) = \begin{cases} \omega - p_i & \text{if } p_i \in [\bar{\omega}_0, \bar{\omega}_1]; \\ 0 & \text{otherwise.} \end{cases}$$

Notice that Equations (3.2.3)-(3.2.4) imply

$$0 < \omega_N < \bar{\omega}_0 < \bar{\omega}_1 < \omega_Y < 1. \quad (3.2.7)$$

In state  $Y$ , when trade occurs, signal-1 agents pay less than  $\bar{\omega}_1$  and therefore less than the settlement value  $\omega_Y$ . They make a profit while sellers (signal-0 agents) sell the asset at a price too low. The opposite applies in state  $N$ . Hence, the center, seeing that people endorsing  $Y$  make a profit, can conclude that we are indeed in state  $Y$ , even though the state itself is not directly observable. Sellers making a profit indicates state  $N$ .  $\square$

In this infinite case, at the truth-telling equilibrium and under the actual state of nature  $S$ , the average payoff for agents with the same signal  $s$  is equal to the expected payoff for agents with that signal:

$$\begin{aligned} \pi_1^Y &\equiv E_p[\pi_1(v_i, p) | Y] \\ &= E_p[\pi_1(\omega_Y, p)] = \int_{\bar{\omega}_0}^{\bar{\omega}_1} (\omega_Y - p) dp = \frac{1}{2} [(\omega_Y - \bar{\omega}_0)^2 - (\omega_Y - \bar{\omega}_1)^2] \\ &= -\pi_0^Y \equiv -E_p[\pi_0(v_i, p) | Y]; \end{aligned} \quad (3.2.8)$$

$$\begin{aligned} \pi_0^N &\equiv E_p[\pi_0(v_i, p) | N] \\ &= E_p[\pi_0(\omega_N, p)] = \int_{\bar{\omega}_0}^{\bar{\omega}_1} (p - \omega_N) dp = \frac{1}{2} [(\bar{\omega}_1 - \omega_N)^2 - (\bar{\omega}_0 - \omega_N)^2] \\ &= -\pi_1^N \equiv -E_p[\pi_1(v_i, p) | N]. \end{aligned} \quad (3.2.9)$$

Under state  $Y$ ,  $\pi_1^Y > 0$  and  $\pi_0^Y < 0$ ; and under state  $N$ ,  $\pi_1^N < 0$  and  $\pi_0^N > 0$ .

The value  $\bar{\omega}_s$  is the prediction of the proportion of signal 1 in the population by agents with signal  $s$ . Hence,  $\omega_s - \bar{\omega}_s$  is the prediction error of signal- $s$  agents when  $S$  is the actual state of nature (note that this error can be positive or negative). The average payoff of signal- $s$  agents are therefore half the difference between the squared prediction error of agents with signal  $1 - s$  and their own squared prediction error. Agents endorsing the actual state of nature are better able to guess the signal distribution in the population, and therefore, the opinions of others. Bayesian markets favor them and allow them to make a profit.

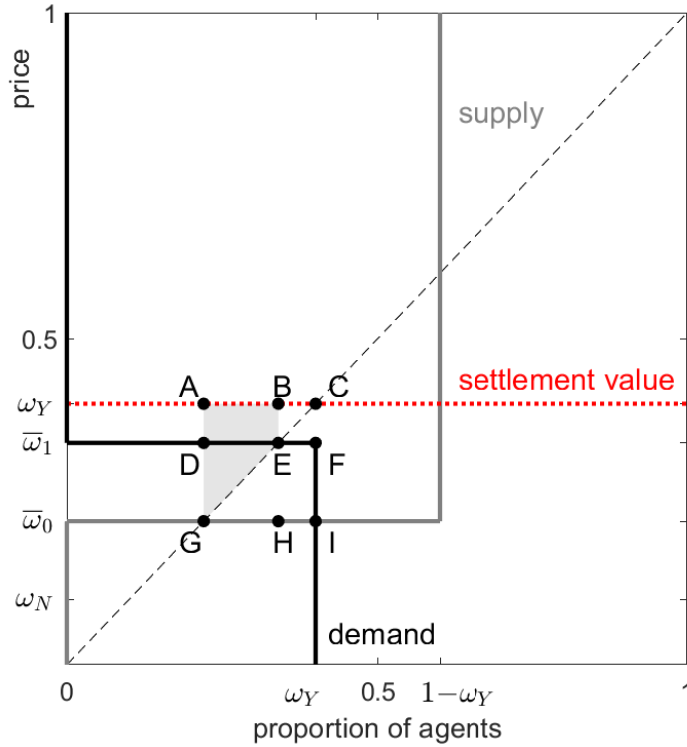


Figure 3.2.1: Demand and supply curves (thick black and gray lines), settlement value (dotted red line), and average payoff of signal-1 agents (shaded area) in state  $Y$  when the majority endorses  $N$ .

Figure 3.2.1 displays the supply and demand curve of the Bayesian market when  $Y$  is the actual state and when only a minority of agents ( $\omega_Y < 0.5$ ) endorsed it. Supply

exceeds demand at prices between  $\bar{\omega}_0$  and  $\bar{\omega}_1$ . The settlement value  $\omega_Y$  is also displayed in the graph. Equation (3.2.8) applies. The squared prediction error of signal-0 agents is represented by square ACIG, and the squared prediction error of signal-1 agents by the smaller square BCFE. The average payoff of signal-1 agents,  $\pi_1^Y$ , hence corresponds to the shaded area (half the difference between ACIG and BCFE).

We can also provide an interpretation of the graph in terms of surplus. Because prices are individual and uniformly distributed, the triangle DEG is both the expected and the average ex ante surplus of buyers (the difference between their willingness to pay and the price they pay). The (expected and average) ex ante surplus of sellers is GEH. The total ex ante surplus of trading is DEHG and is shared equally between buyers and sellers on the Bayesian market. The payoff of an agent is the sum of the ex ante surplus and of the prediction error. For instance, a buyer with  $p_i = \bar{\omega}_0$  gets the maximum ex ante surplus DG, plus the prediction error AD. For buyers, prediction errors are lucky surprises if the state is  $Y$ . The asset is worth more than they expected. A seller with  $p_i = \bar{\omega}_1$  gets the maximum ex ante surplus (EH) but the prediction error BH is an unlucky surprise in state  $Y$ . The asset is more costly than expected. The polygon ABEG represents the sum of the ex ante surplus and of the prediction errors. In state  $Y$ , both terms are positive for buyers but only the former is for sellers. The state  $N$  is entirely symmetric, with all statements reversed for buyers and sellers.

### 3.2.4 Adjustments to a finite population

We presented the results with  $n$  infinite. However, with a small group of agents, the individualized prices may affect the equilibrium strategies. For a buyer, the existence of an agent being offered to sell at a specific price is informative about the number of sellers, and therefore informative about the asset value. Agents may then accept to buy or sell at prices that do not give rise to the aggregation property pointed out by Proposition 3.



Intuitively, the information contained in the existence of an agent on the opposite side of the market should decrease with the number of agents. Below, we make this intuition rigorous: a “sufficiently large” group of agents restores properties used by Propositions 2 and 3. Specifically, we show that for any signal technology there is a truth-telling equilibrium in which agents endorsing the actual state make a positive payoff with arbitrarily large probability, given that  $n$  is larger than some finite threshold  $n^*$ . Thus, as the number of market participants increases, the center can be almost certain that the agents making money are the ones who correctly identify the actual state.

**Proposition 4.** *Let  $\langle \omega_Y, \omega_N, r \rangle$  be a signal technology. Then: (i) truth-telling is an equilibrium in the game induced by the Bayesian Market, and (ii) in this equilibrium, for any  $\varepsilon > 0$ , there exists  $n^*$  such that for all  $n \geq n^*$ , the probability that at least one agent has a positive payoff, and all those with positive payoffs have endorsed the actual state, and all those with negative payoffs have endorsed the opposite state, is at least  $1 - \varepsilon$ .*

The proof is in Section 3.7.2. The biggest obstacle for observing the side of the market who endorsed the actual state is the possibility of no trade. This is especially so in cases in which the proportion  $\omega$  is near the extremes 0 and 1, so that there are only few individuals of one of the types, making it possible that none of them observe a price acceptable to them. There are two main ways to address this issue. First, any information about the signal technology (for instance a lower bound on  $\omega_N$  or an upper bound on  $\omega_Y$ ) would allow the center to draw prices in a more specific region instead of uniformly between 0 and 1, thereby increasing the probability of trade. Second, the absence of trades does not mean that no information was collected. From the agents’ endorsements and trade decisions, the center can recover reservation prices and use the surplus analysis described in Figure 3.2.1 to infer the actual state. The next subsections describes how.

### 3.2.5 Algorithms for empirical data

Propositions 3 and 4 concerns limit behavior of perfectly rational agents. In perfect conditions, all agents endorsing the actual (opposite) state have a nonnegative (nonpositive) payoff, and at least one agent will have a positive payoff. In practical implementation, a small group may lead to no trade. Furthermore, agents may make mistakes when endorsing a state or when deciding to trade, making the supply and demand curves less “sharp” than those depicted in Figure 3.2.1. In the presence of noise, agents not endorsing the actual state may still make a profit. We propose two *follow-the-money* (*FTM*) algorithms which can be used empirically by the center to find the actual state in those non-ideal situations.

The simpler algorithm computes the payoff of each agent and compares the average payoff of the sellers to that of the buyers. The algorithm picks the side with the higher average payoff and tosses a coin if no trade occurred. We call this algorithm *FTM-A* (for average). *FTM-A* is able to accommodate some moderate noise in agents’ behavior but does not solve the no-trade issue.

To account for noise and for no-trade situations, we propose a more elaborate algorithm, fitting logistic curves, called *FTM-L*. With  $F$  the logistic function,<sup>6</sup> *FTM-L* first estimates  $\hat{\omega}_1$  and  $\hat{\omega}_0$  (which can be interpreted as the reservation prices for an infinite group at the truth-telling equilibrium) from

$$Prob(d_i = 1 | p, e_i) = \begin{cases} F(\beta(p - \hat{\omega}_1)) & \text{if } e_i = 1 \\ F(\beta(\hat{\omega}_0 - p)) & \text{if } e_i = 0 \end{cases} \quad (3.2.10)$$

imposing  $\hat{\omega}_0 \leq \hat{\omega}_1$ . Parameter  $\beta$  captures the level of noise/imprecision and is assumed to be the same for sellers and buyers (for parsimony). *FTM-L* then computes the expected

---

<sup>6</sup>If  $F$  is the probit function, we can define another algorithm called *FTM-P*. All the reported results are robust to this specification. We will focus on *FTM-L* from now on.

payoffs for buyers and sellers for an infinite group using Equations (3.2.8) or (3.2.9), substituting  $\bar{\omega}_1$  and  $\bar{\omega}_0$  with estimated reservation prices  $\hat{\omega}_1$  and  $\hat{\omega}_0$ , and  $\omega$  with the proportion of endorsements 1, and picks the side with a positive expected payoff.

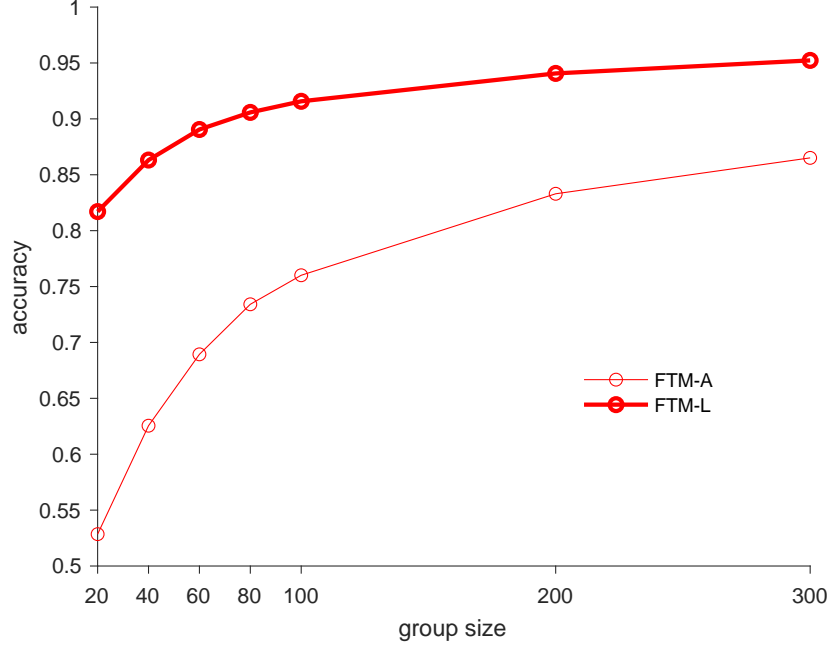


Figure 3.2.2: Simulation results for FTM-A and FTM-L under different group sizes.

Figure 3.2.2 displays the accuracy, defined as the number of times the actual state is selected, of both algorithms in the same simulations as described in the preceding subsection. The simulation does not include noisy agents and therefore, FTM-A cannot improve much with respect to requiring that all agents make a profit to trust their endorsement. By contrast, FTM-L reaches an 80% accuracy with groups as small as 20 agents. It reaches a 95% accuracy with 300 participants, which is not an unusual sample size in social sciences. The cases in which FTM-L is inaccurate correspond to signal technologies that are only very weakly informative. In all other cases, FTM-L recovers all necessary information even in the absence of trades.

### 3.3 Experimental design

#### 3.3.1 Stimuli

We conducted an experiment with abstract tasks (urns and balls) ensuring that the theoretical assumptions were satisfied. We considered groups of  $n = 100$  agents. In each task, the participants of the experiment were presented with two urns, as depicted in Figure 3.3.1. Urns Left and Right represent the two states of nature,  $N$  and  $Y$  respectively. Participants were told that one of the two urns was selected randomly ( $r = 0.5$ ) and that each of the 100 participants of a group would get one ball from that urn. Denoting a yellow ball  $s_i = 1$  and a blue ball  $s_i = 0$ , Urn Left would give  $\omega_N = 0.10$  and Urn Right  $\omega_Y = 0.40$  in this particular example. Urn Right always contains more yellow balls than Urn Left. Thus Urn Right is state of nature  $Y$  and Urn Left is state of nature  $N$ .

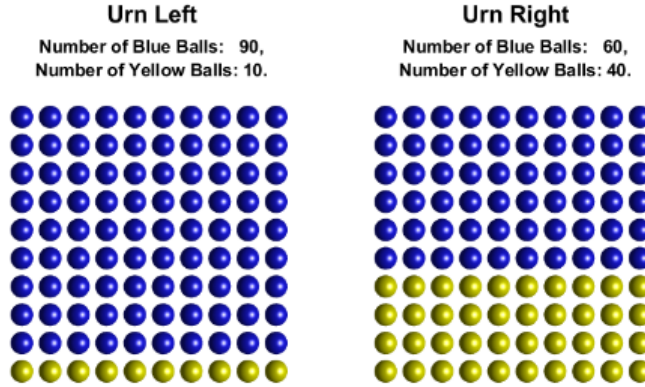


Figure 3.3.1: Experimental task setting (an example of Task 6)

There were 30 tasks with  $\omega_N$  ranging from 0.05 to 0.75 and  $\omega_Y$  from 0.25 to 0.95, spanning the unit interval in a systematic way. In twelve tasks, both urns had a minority of yellow balls, i.e.,  $\omega_N < \omega_Y < 0.5$ . Another set of twelve tasks mirrored them such that  $\omega_Y > \omega_N > 0.5$ , and in six tasks the majority would always guess the correct state of nature ( $\omega_Y > 0.5 > \omega_N$ ). Table 3.7.1 in Section 3.7.3 lists all the task parameters. The

number of yellow balls differs across states of nature by a minimum of 20 and a maximum of 30. Larger differences would mean that the signal technology discriminates very well between state of nature and the majority (as well as FTM) would be right most of the time. By contrast, smaller differences would imply very narrow trading intervals  $[\bar{\omega}_0, \bar{\omega}_1]$  and it could be that none of the 100 participants of a group gets a price in that range.

In each task, the participants were first presented with the urns (as in Figure 3.3.1) and asked to press a button to draw their ball. Once the color of their ball was revealed, they were asked to guess which urn the ball comes from (i.e. to endorse a state). The next question differed between two experimental treatments, FTM and SPA.

Your draw:



You guessed that your draw came from **Urn Right**.

Do you bet that the number of participants guessing **Urn Right** is at least **41**?

Yes

No

Figure 3.3.2: Screenshot of the FTM treatment

In the FTM treatment, we implemented the betting mechanism of the Bayesian markets. In Figure 3.3.2 for instance, participants were asked whether they were willing to bet that the number of participants guessing Urn Right (i.e. endorsing  $Y$ ) was at least 41, i.e. whether they were willing to pay  $p = 0.41$  for the sample proportion  $\omega$ . For the sake of symmetry, participants guessing Urn Left were asked whether they would bet that the number of participants guessing Urn Left would be at least 59, i.e. whether they were willing to accept  $p = 0.41$  for  $v$ . Payment was explained in a training preceding the exper-

iment. The participants were told that the number (e.g. 41) was random and that their payment would be the actual number of Urn Right guesses minus that number if someone took the opposite bet (betting that at least 59 participants would guess Urn Left). It would be 0 otherwise.

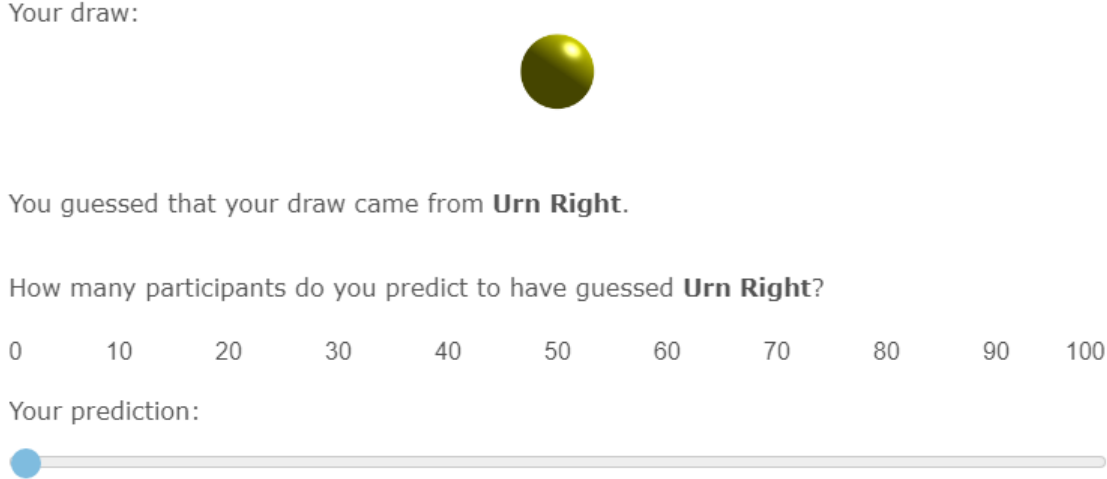


Figure 3.3.3: Screenshot of SPA.

In the SPA treatment, we followed the approach of Prelec et al. (2017) and asked participants to predict the number of people who guessed the same urn as they did (Figure 3.3.3). Prelec et al.'s (2017) algorithm first computes the average prediction across all participants and then selects the state of nature that was endorsed more often than predicted. Predictions were incentivized using the quadratic scoring rule. Participants received  $200 - \frac{x^2}{50}$ , with  $x$  the difference between their prediction and the actual number of guesses.

In none of the treatments was payment directly based on the task parameters. Even though we, the experimenters, knew them, we aimed to mimic situations in which no one knows the actual state of nature and in which the center (paying the agents) does not even know the signal technology.

### 3.3.2 Deviations from Section 3.2.2

The implementation of the Bayesian market in our experiment differs from the Bayesian market mechanism proposed in Subsection 3.2.2 in two ways. First, the settlement value  $v$  did not exclude the agent’s own endorsement. It allowed us to present all relevant values as shares of 100 (and not 99 or 101). Furthermore, the draws from the urn (i.e., the signals) were made without replacement. As a consequence, the settlement value could only be  $\omega_Y$  or  $\omega_N$ , as in the infinite case.

### 3.3.3 Implementation

The experiment was conducted on Prolific in August 2019, with 473 participants in the FTM treatment and 462 in the SPA treatment. They were all US students. We restricted participation to students for their probable familiarity with abstract tasks as those used in our experiment. Participants watched a short video explaining the experimental tasks and then went through five training rounds where they received feedback about their payments and how these payments were calculated (see Section 3.9 for details). We split the 30 tasks into two sets of 15. After the training, each participant completed one of the two sets, with the task order being randomized within that set at the participant level. There was no feedback after the tasks. Payment, described in the next paragraph, occurred once all participants had completed the experiment.

Participants received a fixed reward of £1.5 and a bonus of up to £3.<sup>7</sup> All amounts (prices, bets, scores) were presented in tokens. The bonus in pounds was the number of tokens divided by 1,000. In the FTM treatment, participants could (in theory) win or lose up to 100 tokens in each task. Hence, they were endowed with 100 tokens for each task to avoid net losses at the end of the experiment. In the SPA treatment, the quadratic score was also expressed in tokens. It was equivalent to endowing them with 200 tokens and

---

<sup>7</sup>Prolific requires payments in pounds.

imposing a quadratic loss ranging from 0 to 200. In both treatments, the final number of tokens was naturally bounded by 0 and 3000. This allowed us to recruit participants with the same information about bonus ranges. However, the average bonus was likely to be lower for the FTM treatment than for the SPA treatment *ex ante* and, in fact, it was *ex post* (SPA £2.85, FTM £1.60).

To compute the bonus of a participant in a given task after the end of the experiment, we randomly selected a state of nature<sup>8</sup> and 100 participants such that the group (including this particular participant) had the exact combination of signals shown in the task. In other words, participants were not assigned to a given group *ex ante*. Instead, we constructed (random) groups matching the information provided to the participants.

## 3.4 Results

To be consistent, we report data and results in terms of our theoretical setting. In particular, a yellow ball is signal 1 ( $s_i = 1$ ) and a blue ball is signal 0 ( $s_i = 0$ ). A participant guessing Urn Right is endorsing state of nature  $Y$  ( $e_i = 1$ ) and guessing Urn Left is endorsing  $N$  ( $e_i = 0$ ).<sup>9</sup> We also define truth-telling as reporting  $e_i = s_i$ .

### 3.4.1 Raw data - Endorsements

According to the model, truth-telling is a Bayesian Nash equilibrium in the FTM treatment. The empirical truth-telling rate was 87.6%.<sup>10</sup> About 53% of the participants told

---

<sup>8</sup>This random selection of a state of nature resulted in 50.2% of state  $Y$  selected for bonus calculations of participants in the SPA treatment, and 48.9% for the FTM treatment. Both proportions are not significantly different from 0.5 (proportion tests: for SPA,  $Z$ -statistic= 0.333,  $p = 0.739$ ; and for FTM,  $Z$ -statistic= 1.904,  $p = 0.057$ ).

<sup>9</sup>Predictions elicited in the SPA treatment were about the number of participants guessing the same urn, but we deduct the predictions of participants endorsing  $N$  from 100 to be the predictions of number of participants endorsing  $Y$ . Bets in the FTM treatment were also expressed in terms of the number of participants guessing the same urn, but we deduct the prices in the bets for participants endorsing  $N$  from 100 to be the prices to sell the asset whose settlement value is the number of participant endorsing  $Y$ .

<sup>10</sup>The empirical truth-telling rates were not significantly different for easier questions with  $\omega_Y > 0.5 > \omega_N$  and for other questions with  $\omega_Y > \omega_N > 0.5$  or  $0.5 > \omega_Y > \omega_N$  (87.0% and 87.7% respectively; pro-



the truth in all 15 tasks they faced. About 25% guessed the opposite urn (or lied about their guess) 1 to 3 times out of 15. Less than 4% had a majority of lies / wrong guesses (Table 3.7.2 in Section 3.7.4). The incentives provided in the SPA treatment did not make truth-telling a Bayesian Nash equilibrium, but we observed a very similar truth-telling rate (SPA: 87.8% of the cases, not significantly different from FTM, with proportion test  $Z$ -statistic= 0.376 and  $p = 0.704$ ).

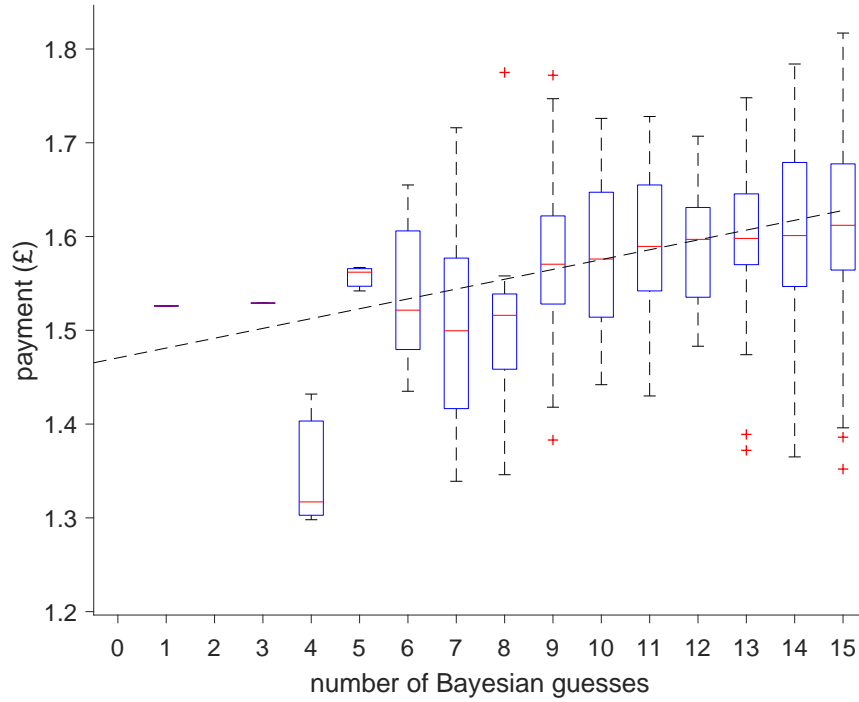


Figure 3.4.1: Number of Bayesian guesses and payment in FTM.

The truth-telling rate of the FTM treatment was sufficiently high to reward those who correctly reported their signal and to penalize those who lied or misreported. Figure 3.4.1 displays earnings as a function of the number of times people told the truth. It shows a positive correlation, with a fitted line slope of 0.010 ( $p < 0.001$ ). People did not get feedback during the experiment (only in the five training rounds). Figure 3.4.1 illustrates that feedback about payment could have improved truth-telling rate by allowing

---

portion test  $Z$ -statistic= 0.755 and  $p = 0.450$ ).

participants to learn that correctly reporting their signal is rewarded. It further shows that in future experiments one can announce in the instructions that a previous study showed that participants who tell the truth more often can earn more in such a setting.

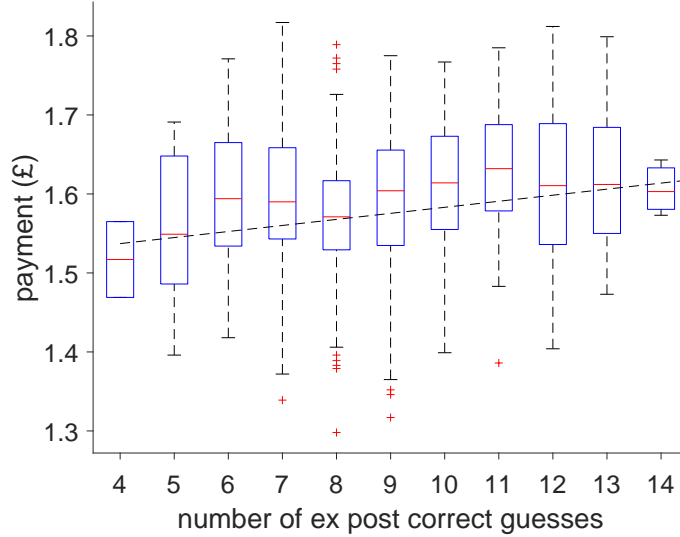


Figure 3.4.2: Number of ex-post correct guesses and payment in FTM.

So far, we studied what the raw data told us about participants' strategic behavior, illustrating the incentive properties of Bayesian markets (Proposition 2). To illustrate the aggregation properties (Proposition 3), we can check whether correctly guessing the selected urn led to higher earnings in our experiment. The prediction is supported by Figure 3.4.2, which is a box plot of earnings as a function of the number of times participants guessed the actual state. The fitted line slope is 0.008 ( $p < 0.001$ ). Thus, Bayesian markets reward expertise. While in our experiment, this expertise is artificially created,<sup>11</sup> in many applications one may expect that the number of times someone guesses the actual state of the world to be influenced by a more natural notion of expertise, i.e. domain knowledge.

<sup>11</sup>It consists of receiving informative signals, in combination with a truth-telling strategy.

### 3.4.2 Raw data - Predictions and trades

If participants are Bayesian, they should report the posteriors  $\bar{\omega}_0$  and  $\bar{\omega}_1$  in the SPA treatment, at least if they expect everyone else to tell the truth. Figure 3.4.3 displays the average predictions as a function of theoretical posteriors for both type of guess. Predictions are very close to Bayesianism for  $\bar{\omega}_0 < 0.5$  and  $\bar{\omega}_1 > 0.5$ . Interestingly, participants seemed to have much more difficulty to predict that a majority of people would guess  $Y$  when they themselves guess  $N$  or that only a minority would guess  $Y$  when they themselves guess  $Y$ . The SPA uses the average prediction across both guesses, which mitigates this issue.

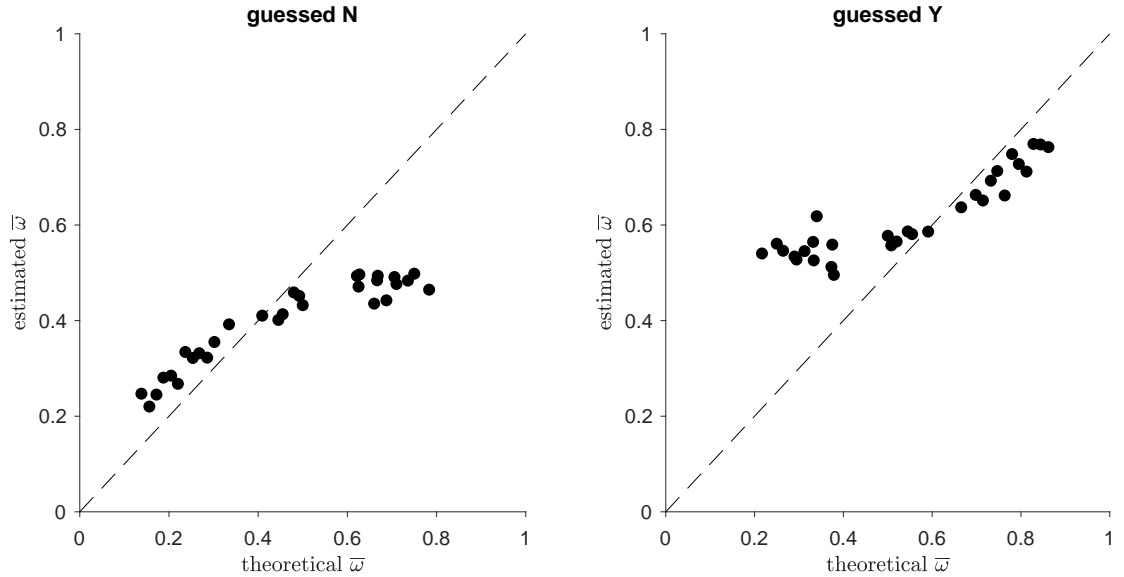


Figure 3.4.3: Theoretical  $\bar{\omega}_0$  and  $\bar{\omega}_1$  vs. average predictions in SPA.

We do not have people's predictions in the FTM treatment but we can compare the participants' decisions  $d_i$  to the theoretical predictions. Table 3.4.1 compares the theoretical and empirical proportions of  $d_i = 1$  (the willingness to buy / to sell) for five price intervals, defined with  $\omega_N$ ,  $\bar{\omega}_0$ ,  $\bar{\omega}_1$ , and  $\omega_Y$ . Buyers should be willing to pay at most  $\bar{\omega}_1$  and sellers willing to accept not less than  $\bar{\omega}_0$ . If participants do not compute the Bayesian

posterior but use  $\omega_Y$  and  $\omega_N$  instead, i.e. the distribution of balls of the urn they guessed, buyers would be willing to pay at most  $\omega_Y$  and sellers willing to accept not less than  $\omega_N$ . If they were extremely risk averse, buyers would be willing to pay at most  $\omega_N$  and sellers willing to accept not less than  $\omega_Y$ .

The empirical willingness to sell was increasing with price and the empirical willingness to buy was decreasing, as predicted in the truth-telling equilibrium. However, for several participants the acceptance and rejection ranges of bets were not consistent with the equilibrium prediction. About 30% of bets that are losing for sure under truth-telling were accepted and about 10% of bets that are winning for sure under truth-telling were rejected (see leftmost column of the seller row and rightmost column of the buyer row). In total, there was a clear tendency to bet much more than predicted by equilibrium play.

Table 3.4.1 also reports the average payoffs of the participants for each price interval. The results confirm that participants who accepted bets that would have been losing for sure if everyone else had told the truth, still bore a loss on average in our experiment. Overall, trading decisions were noisy and substantially deviated from the theoretical predictions. This underlines that the performance of the FTM algorithms will depend on their ability to recover aggregate reservation prices from the noisy trades.

Table 3.4.1: Theoretical and empirical bet acceptance (in %) and average payoffs (in tokens) in the FTM treatment by price interval

$p \in$		$[0, \omega_N)$	$[\omega_N, \bar{\omega}_0)$	$[\bar{\omega}_0, \bar{\omega}_1]$	$(\bar{\omega}_1, \omega_Y]$	$(\omega_Y, 1]$
guessed $N$ (seller)	theo. acceptance	0%	0%	100%	100%	100%
	emp. acceptance	30.4%	55.5%	67.6%	80.8%	90.6%
	average payoff	-27.2	-6.1	2.2	10.8	19.3
guessed $Y$ (buyer)	theo. acceptance	100%	100%	100%	0%	0%
	emp. acceptance	89.9%	80.3%	67.7%	48.4%	29.1%
	average payoff	21.0	9.9	1.3	-6.9	-25.0

### 3.4.3 Accuracy comparison

The final part of the analysis aims to compare accuracy of the various methods. We want to assess the ability of the majority rule, SPA, and FTM algorithms to identify the actual state of nature using the participants' answers.

To make full use of the answers of all respondents who provided answers to a task, we ran 1,000 simulations for each task, state of nature, and treatment, randomly making groups of 100 participants. For instance, consider one of the simulations for the task described in Figure 3.3.1 with  $\omega_Y = 0.40$  and  $\omega_N = 0.10$  (Task 6), state of nature  $Y$ , and the FTM treatment. We randomly composed a group of 100 FTM participants, such that exactly 40 of them had gotten  $s_i = 1$ . We then use the answers from the 100 participants to determine the state using majority rule, FTM-A, and FTM-L. Figure 3.4.4 shows the results from the FTM-L algorithm by fitting Equation (3.2.10). It is the empirical counterpart of Figure 3.2.1. The estimated  $\hat{\omega}_0$  and  $\hat{\omega}_1$  are more extreme than the theoretical  $\bar{\omega}_0$  and  $\bar{\omega}_1$ , and visually indistinguishable from  $\omega_N$  and  $\omega_Y$  respectively.<sup>12</sup> Once the curves displayed in the figure are obtained, the FTM-L algorithm uses the sample proportion  $v$ ,  $\hat{\omega}_0$ , and  $\hat{\omega}_1$  to determine the expected payoffs of the representative buyers and sellers. In the case described in Figure 3.4.4, FTM-L could still detect the actual state  $Y$ .

Similarly, we randomly composed 1,000 groups of 100 SPA participants in the same way to determine the state using majority rule and SPA. Repeating the same procedures for each of the 30 tasks and two possible states of nature, we obtained 60 accuracy rates for each method. Table 3.4.2 summarizes the average accuracy rates for each algorithm and for the majority rule. We conducted Wilcoxon tests to test for differences.

Table 3.4.2 distinguishes two cases. If  $\omega_N < 0.5 < \omega_Y$  (top row), then the majority rule should determine the actual state all the time. In the other cases (bottom row), the

---

<sup>12</sup>Note that this observation is not generally applicable to all the 1000 simulations.

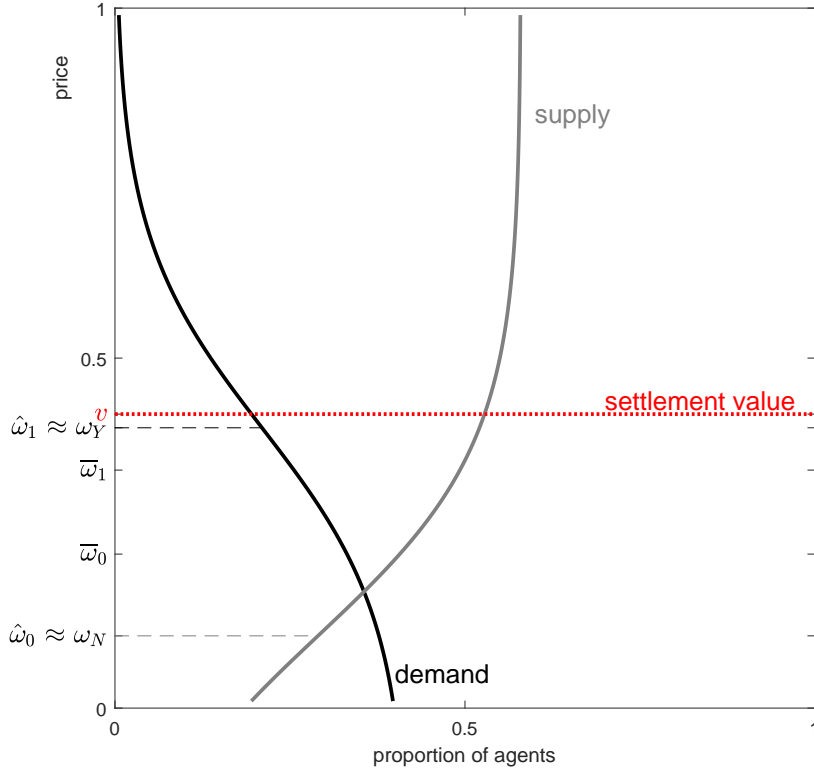


Figure 3.4.4: An example of a simulation in Task 6 under state Y.

Table 3.4.2: Average accuracy rates from simulations

cluster of questions	majority rule		SPA	FTM	
	data SPA	data FTM		<i>FTM-A</i>	<i>FTM-L</i>
$\omega_Y > 0.5 > \omega_N$	95.7%	99.7%	93.2%	87.3%	91.5%
$\omega_Y > \omega_N > 0.5$ or $0.5 > \omega_Y > \omega_N$	51.0%	53.8%	73.9%	62.8%	75.0%

majority rule finds the actual state 50% of the time, by pure chance. Our results are consistent with these predictions (see columns ‘majority rule’), both for the data from the SPA treatment and for the data of the FTM treatment. The SPA, our benchmark, should always identify the actual state if participants were Bayesian and reporting truthfully all the time. Non-Bayesian answers and noise make the SPA perform worse (Wilcoxon signed rank test  $p = 0.031$ ) than majority when  $\omega_N < 0.5 < \omega_Y$  but substantially improved upon majority when following the majority is equivalent to tossing a coin (Wilcoxon signed rank

test  $p < 0.001$ ). In that case, the average accuracy increased by 22.9 percentage points (pp).

Computing average payoffs on Bayesian markets, as our FTM-A algorithm does, led to worse results than SPA, whether  $\omega_N < 0.5 < \omega_Y$  (Wilcoxon signed rank test  $p = 0.054$ ) or not (Wilcoxon signed rank test  $p = 0.001$ ). FTM-A is highly sensitive to noise and we noticed earlier that our data were clearly noisy. To account for noise, FTM-L fits logistic supply and demand curves on the buy and sell decisions and only then computes expected payoffs. FTM-L substantially improved upon FTM-A (Wilcoxon signed rank test  $p < 0.001$ ), especially when  $\omega_Y > \omega_N > 0.5$  or  $0.5 > \omega_Y > \omega_N$ , with an increase of 12.2pp. It yielded results that were not significantly different from SPA (Wilcoxon signed rank tests; top row  $p = 0.322$  and bottom row  $p = 0.845$ ). Interestingly, it gave results comparable to SPA with less information. SPA uses, as input, an endorsement and a prediction (number between 0 and 1), directly asking participants for  $\bar{\omega}_0$  and  $\bar{\omega}_1$ . FTM-L uses an endorsement and a trade decision, which is binary. FTM-L compensates the information loss by using (simple) econometric techniques to recover reservation prices, which should be  $\bar{\omega}_0$  and  $\bar{\omega}_1$  at the truth-telling equilibrium.

The accuracy analysis so far was based on groups of 100 participants. We can also study how sensitive results are to group sizes. We replicated the analysis (with 1,000 simulations for each combination of method, task, and state of nature) for various group sizes ranging from 20 to 100. Figure 3.4.5 depicts the accuracy rates as a function of group size. FTM-L is more sensitive to group size than SPA. In the left panel, when  $\omega_N < 0.5 < \omega_Y$ , accuracy of FTM-L increases from around 75% for groups of 20 to more than 90% for groups of 100. SPA performs better for small groups but accuracy increases less with group size. In the right panel ( $\omega_Y > \omega_N > 0.5$  or  $0.5 > \omega_Y > \omega_N$ ), SPA is very stable, with accuracy rates between 72% and 74%. The accuracy of FTM-L is lower than that of SPA for groups of 20 but slightly higher for groups of 100. These results are not surprising, knowing that, for groups of 20, FTM-L has to determine reservations prices

from very few binary decisions (buying or selling).<sup>13</sup>

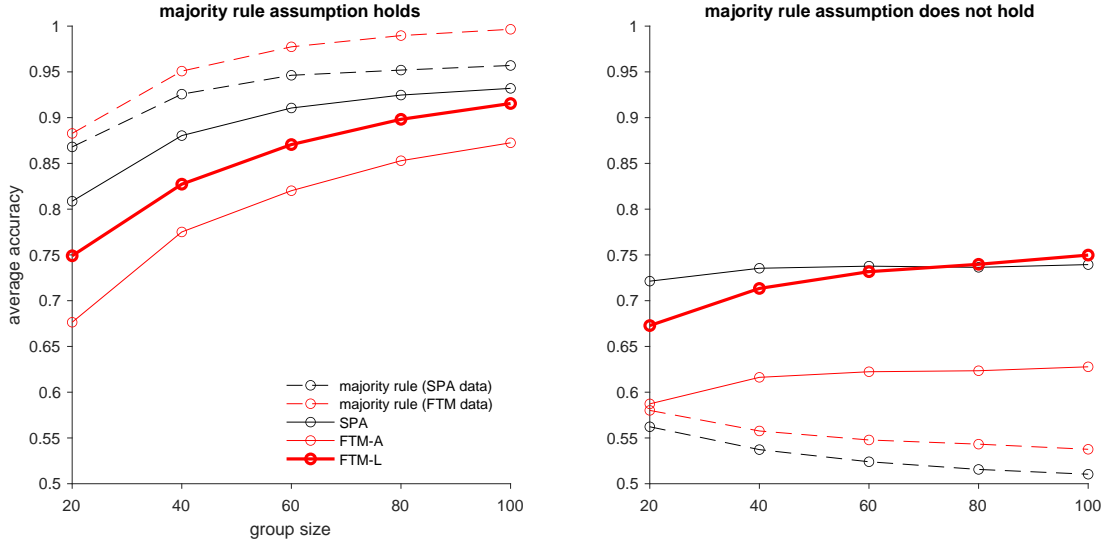


Figure 3.4.5: Accuracy comparisons for different group sizes.

### 3.5 Discussion

In our experiment, each agent received an endowment to avoid losses. Even without providing an endowment, agents can expect a strictly positive payoff (Baillon, 2017), which can motivate them to participate. The center, who plays the role of the market maker, subsidizes the market and acts as an intermediary between the agents, who do not trade with each other. Absent this intermediary role of the center, agents would infer others' signals from their willingness to buy or sell. Similar to the classical reasoning in Aumann (1976) and Milgrom and Stokey (1982), they would then agree on the state, leaving no room for trade based on disagreement.

To communicate the same information to all potential participants, we fixed the bonus range from £0 to £3. The SPA was more expensive (SPA £2.85, FTM £1.60). If anything, the SPA participants, with an endowment of 200 tokens and a quadratic loss should have

<sup>13</sup>There may be as little as one buyer or one seller in some groups even if everybody reports truthfully.



been more motivated than the FTM participants. The SPA treatment only incentivized predictions, not truthful endorsement. The latter could have been done using the Bayesian truth serum of Prelec (2004) but the payoff rule is difficult to explain to participants. Experiments that have been using this truth serum did not explain the payoff function in detail, but rather used an “intimidation method”, telling participants it is in their interest to tell the truth. We refrained from such an approach, and instead included instructions and training to explain our payoff rules. An alternative for future research is to incentivize the SPA using choice-matching, which will be discussed in the next chapter. Choice-matching elicits predictions and endorsements with a simpler payment formula.

Regarding our theoretical model, four restrictive assumptions warrant some further discussion. First, recall that that we treated strategies as maps from signals to endorsements, such that agents could not make their endorsements depend on any other event or randomization device, and did not allow asymmetric strategies. Second, we allowed no communication between agents. Third, we only considered a binary underlying state space and fourth, our market setting is a one-period, static setting. We discuss each of those in turn.

It is important that agents cannot coordinate on events other than type realizations. Among the remaining symmetric equilibria, the truth-telling equilibrium is (ex-ante) Pareto optimal.<sup>14</sup> It is behaviorally plausible, as conjectured by Baillon (2017), that truth-telling is focal and, in our experiment, there was indeed little evidence of agents trying to find a reverse strategy. More than half of the participants consistently told the truth and a negligible share of participants chose to systematically misstate their type (see Table 3.7.2 in Section 3.7.4). Without the aforementioned restriction, agents could try to coordinate on some other signal, in which the probability of receiving a 0-signal in state  $Y$

---

<sup>14</sup>To see this, note first that it is pay-off equivalent to the “reverse” equilibrium in which everyone endorses the state that they believe to be less likely to be the actual one. The two other equilibria have universal endorsement of either state  $Y$  or of state  $N$  and thus obviously lead to a universal payoff of zero since there is never any trade.

and a 1-signal in state  $N$  is very low. Then, a small number of agents will make a loss of (almost) 1, and a large share of agents will make a profit of (almost) 1. In expectation, all agents thus have a high expected payoff. Note that this coordination does not only require mere communication among respondents but also some credible randomization device. To avoid such coordinated attacks, the center should make it an active feature of design that market participants are (at least partially) anonymous, as is the case in our experiment.

As suggested by the previous paragraph, our approach cannot be used if there is public discussion of private signals or if agents can form coalitions. If it is possible to bring all experts together, other approaches to the aggregation problem have been proposed in the literature, such as the Delphi method developed in the 1950s at the Rand Corporation (Okoli and Pawlowski, 2004). These approaches do not solve the incentive problem though. In our setting, experts do not have other incentives than those we provide to hide or manipulate their private information. The literature on committee decisions studies how agents may agree to share their private signals with each other in order to look united if their reputation is at stake (Visser and Swank, 2007; Swank et al., 2008).

We considered a binary state space. If the state space is non-binary, one may organize several Bayesian markets, with different agents. Consider three states  $A$ ,  $B$ , and  $C$ , and assume agents can choose which state they would like a signal about (e.g., an agent can design an experiment testing whether we are in state  $A$  or not- $A$ ). The center can assign agents to markets, inform them about which state their market will be about, let them run their experiment for that state, and then organize the Bayesian markets.

Bayesian markets and their aggregation properties can further be translated to a setting in which a market is run continuously. Suppose that there are  $T$  periods and that for each  $t = 1, \dots, T$ , a Bayesian market is set up to trade on an asset  $v_t$  that represents the share of buyers in the Bayesian market at time  $t$ . All of these markets are only settled at the final period  $T$ , so that in particular agents do not learn the value of the assets. At each  $t$ , the incentive and aggregation properties of Bayesian markets are not affected by the

markets in other periods. A continuous market can sometimes be advantageous for the center: Suppose for instance that the signal technology is constant across all periods, but that the actual state  $S$  (and therefore  $\omega_S$ ) may vary with  $t$ . Once the center has found a market-clearing price  $p^*$  (i.e. a price at which each agent is willing to either buy or sell the asset), this price can be chosen for any subsequent period. Since the signal technology is the same, this price will now lead to trade in each period, thereby reducing the payoff-uncertainty faced by the agents. Then, the center can make inferences about the change of the state over time by computing which side would make a profit if the market was settled. Furthermore, if the signal technology is not fixed, this will be reflected in the buying and selling decisions of the agents, and henceforth the center can detect such changes.

The literature on the wisdom of crowds started with the intuition that asking many people may be better than relying on a few experts. Some have raised doubts on the mere possibility to “chase the experts” within a group (Larrick and Soll, 2006). However, there is still value to ask large groups of experts. DellaVigna and Pope (2017) found that the aggregated opinion of academic experts is closer to experimental results than estimates based on a meta-analysis of previous empirical findings. In a follow-up study, DellaVigna and Pope (2018) also showed that academic experts better predict than non-experts, even though degrees of expertise (among experts) such as academic rank or citations do not correlate with performance. Aggregating the opinions of very large group of experts becomes more and more common, for instance the International Panel on Climate Change or surveys of economists and financial specialists about future economic indicators.

Increasingly in companies, prediction markets are used internally among employees to forecast short-term company performance and external events for decision making. Examples include Siemens (Ortner, 1998), Nokia (Hankins and Lee, 2011), Hewlett-Packard (Plott and Chen, 2002), Intel (Gillen et al., 2017), Google (Cowgill et al., 2009), and Ford Motor Company (Cowgill and Zitzewitz, 2015). Results are promising, showing the potential of markets as an effective information aggregation tool in practice. For instance, in the

case of Ford where weekly auto sales forecasts are taken extremely seriously for planning procurement and production, forecasts from the internal prediction markets still outperformed other forecasts available to management (Cowgill and Zitzewitz, 2015). Since Bayesian markets do not require the predicted events to be verifiable in the short-term or at all, companies can potentially use internal Bayesian markets to forecast long-term events or even counterfactual events.

## 3.6 Conclusion

Prediction markets are increasingly used to incentivize and aggregate expert opinions. They are not applicable though if the state of the world is not objectively observable. In such a case, payoffs cannot be state-contingent, creating an incentive problem. Furthermore, in many plausible situations, one may prefer not to rely on the majority opinion, at least if experts themselves, aware of the signal structure, would not. We demonstrated theoretically and empirically how to solve both the incentive and the aggregation problem at once. Agents bet on others' endorsement and their payoffs reveal the state of nature. When implemented in a large online experiment, our follow-the-money approach performed as well as a recent alternative, the surprisingly popular algorithm, but made use of less information from participants. With these results, we hope to add to the promise of prediction markets (Arrow et al., 2008), the promise of Bayesian markets.

### 3.7 Proofs

To simplify the notations in Sections 3.7.1 to 3.7.2, we define  $\omega_{-i} = \frac{\sum_{j \neq i} s_j}{n-1}$ ,  $\tilde{P}_s(\cdot) \equiv P(\cdot | s_i = s, \omega_{-i} \in (0, 1))$  and  $\tilde{E}_s[\cdot] \equiv E[\cdot | s_i = s, \omega_{-i} \in (0, 1)]$ , omitting the conditioning on  $i$ 's signal being  $s$  and the fact that there is at least one signal-1 agent and one signal-0 agent other than  $i$ . Furthermore,  $\tilde{r}_s \equiv \tilde{P}_s(Y)$  and  $\tilde{\omega}_s \equiv \tilde{E}_s[\omega_{-i}]$ . Finally, for any agent  $i$  let  $\mathcal{E}_s^{-i}(p', p'')$  be the event that there exists an agent  $j \neq i$  such that  $s_j = s$  and  $p' \leq p_j \leq p''$ .

All the following proofs establish the results for  $s_i = 1$ . Equivalent proofs for  $s_i = 0$  can be immediately obtained by replacing  $s_i, e_i, \omega_{-i}, p_i$  by  $1 - s_i, 1 - e_i, 1 - \omega_{-i}, 1 - p_i$ . Recall that  $n \geq 4$  for all the proofs.

#### 3.7.1 Lemmas

The following lemmas are used to establish Proposition 4 (Lemma 1 being used to establish Lemma 6). The lemmas are not especially surprising, and the proofs are more cumbersome than truly informative. We therefore relegated them to Section 3.8.

**Lemma 1.**  $\lim_{n \rightarrow \infty} \tilde{\omega}_1 = \bar{\omega}_1$  and  $\lim_{n \rightarrow \infty} \tilde{\omega}_0 = \bar{\omega}_0$ .

**Lemma 2.** For any  $\varepsilon > 0$ , there exists  $n^*$  such that for all  $n \geq n^*$ ,  $P(\omega_{-i} > \bar{\omega}_1 | Y) > 1 - \varepsilon$ ,  $P(\omega_{-i} < \bar{\omega}_0 | N) > 1 - \varepsilon$ ,  $P(\omega_{-i} > \tilde{\omega}_1 | Y) > 1 - \varepsilon$ , and  $P(\omega_{-i} < \tilde{\omega}_0 | N) > 1 - \varepsilon$ .

**Lemma 3.** The expectation  $\tilde{E}_s[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')]$  is increasing in  $p''$  and decreasing in  $p'$ . The expectation  $\tilde{E}_s[\omega_{-i} | \mathcal{E}_1^{-i}(p', p'')]$  is decreasing in  $p''$  and increasing in  $p'$ .

**Lemma 4.** For any prices  $p', p''$ ,  $\tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] < \tilde{\omega}_1 < \tilde{E}_1[\omega_{-i} | \mathcal{E}_1^{-i}(p', p'')] and  $\tilde{E}_0[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] < \tilde{\omega}_0 < \tilde{E}_0[\omega_{-i} | \mathcal{E}_1^{-i}(p', p'')]$ .$

**Lemma 5.**  $\tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')]$  and  $\tilde{E}_0[\omega_{-i} | \mathcal{E}_1^{-i}(p', p'')]$  are continuous in  $p'$  and  $p''$ .

**Lemma 6.** *For any prices  $p', p''$ ,*

$$\lim_{n \rightarrow \infty} \tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] = \bar{\omega}_1$$

*and*

$$\lim_{n \rightarrow \infty} \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p', p'')] = \bar{\omega}_0.$$

**Lemma 7.** *There exists  $n^*$  such that for all  $n \geq n^*$ , there are prices  $p^*, p^{**}$  such that*

$$p^* - \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p^*, p^{**})] = 0$$

*and*

$$\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p^{**})] - p^{**} = 0$$

.

**Lemma 8.** *For any prices  $p', p''$ , there exists  $n^*$  such that for all  $n \geq n^*$ ,  $\frac{\partial}{\partial p}$*

$$\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p', p)] < 1$$

*and*

$$\frac{\partial}{\partial p} \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p, p'')] < 1$$

*for all  $p' < p \leq p''$ .*

### 3.7.2 Proof of Proposition 4

**Proof of (i)** Suppose that all agents are truth-telling, so that  $v_{-i} = \omega_{-i}$ . For  $p', p'' \in [0, 1]$ , consider the expectation  $\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')]$ . From Lemma 7, there are prices  $p^*, p^{**}$  such that  $p^* - \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p^*, p^{**})] = 0$  and  $\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p^{**})] - p^{**} = 0$  for sufficiently large  $n$ . Suppose next that signal-1 agents will buy at any price  $p \leq p^{**}$  and signal-

0 agents will sell at any price  $p \geq p^{**}$ . From Lemma 8, we have  $\frac{\partial}{\partial p} \tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p)] < 1$  and  $\frac{\partial}{\partial p} \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p, p'')] < 1$  for any  $p \in (p^*, p^{**})$  (given sufficiently large  $n$ ). As a result, we have for all  $p < p^{**}$  and for all  $p > p^*$ , respectively:

$$\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p)] - p > 0 \quad (3.7.1)$$

$$p - \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p, p^{**})] > 0. \quad (3.7.2)$$

Given the strategies above, signal-1 agents buy at any price  $p$  which is at most  $\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p)]$  and signal-0 agents will sell at any price  $p$  which is at least  $\tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p, p^{**})]$ . To see that these strategies constitute an equilibrium for sufficiently large  $n$ , we can take the perspective of a buyer. It is completely analogous to show that there is no profitable deviation for a seller. Consider an agent  $i$  who observes signal  $s_i = 1$  and price  $p_i$ . Under the strategy described above,  $i$  never accepts to trade at a price for which  $i$  expects a negative payoff and thus  $i$ 's expected payoff is strictly positive. Any deviation in which  $i$  reports  $e_i = 1$  but buys at a price  $p_i > \tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p_i)]$  is at best non-harmful, if no seller sells at this price, or leads to a lower expected payoff. Suppose then next that  $i$  deviates by reporting  $e_i = 0$ . Afterwards,  $i$  cannot do better than selling at any price  $p_i$  for which  $p_i > \tilde{E}_1 [\omega_{-i} | \mathcal{E}_1^{-i}(p_i, p^{**})]$ . From Lemma 4, we know that  $\tilde{E}_1 [\omega_{-i} | \mathcal{E}_1^{-i}(p_i, p^{**})] > \tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p_i, p^{**})]$ . Hence, all other signal 1-agents reject such trades, and therefore,  $i$ 's payoff will be zero. Thus, there is no profitable deviation for  $i$  and the strategies described above indeed constitute an equilibrium.

**Proof of (ii)** To prove the second part of Proposition 4, we need to find a group size such that we are almost certain that at least one agent has a positive payoff, and all those with positive payoffs have endorsed the actual state, and all those with negative payoffs have endorsed the opposite state. Alternatively, we can show that, for a given group size

and above, the probability of the following three events is negligible:

- Event 1: there exists an agent endorsing the actual state making a loss;
- Event 2: there exists an agent endorsing the other state making a profit;
- Event 3: no agent trades.

Let us first show that Event 2 implies Event 1. In state  $Y$ , an agent endorsing  $e_i = 0$  can only make a profit if there is a buyer  $j$  with  $p_j \geq p_i > v_i$ . Since  $v_i$  excludes  $e_i = 0$  and  $v_j$  excludes  $e_j = 1$ , we have  $v_j = v_i - \frac{1}{n-1} < v_i$ . Then  $p_j > v_j$ , implying that agent  $j$  must be making a loss. The same argument can be made for state  $N$ . Consequently, Event 2 is a subevent of Event 1 and can be ignored.

We can establish  $P(\text{"agents endorsing the actual state, and only them, will make a profit"}) > 1 - \varepsilon$  for all  $n$  above a threshold by showing that  $P(\text{Event 1}) < \varepsilon/2$  and  $P(\text{Event 3}) < \varepsilon/2$ . Recall that statement (ii) posits that we are in a truth-telling equilibrium. We can therefore equate  $v_i$  with  $\omega_{-i}$ .

**Event 1: there exists an agent endorsing the actual state making a loss.**

From Lemma 2, it follows that we can find  $n_1$  such that in the truth-telling equilibrium with probability at most  $\frac{\varepsilon}{2}$ , we have  $v_i \leq \tilde{\omega}_1$  in state  $Y$  or  $v_i \geq \tilde{\omega}_0$  in state  $N$ . Furthermore, from Lemma 4, we have that for any prices  $p', p''$  and any  $n$ :  $\tilde{E}_1[v_i | \mathcal{E}_0^{-i}(p' p'')] < \tilde{\omega}_1$ , and  $\tilde{E}_0[v_i | \mathcal{E}_1^{-i}(p' p'')] > \tilde{\omega}_0$ . Thus, in the equilibrium described in part (i) of Proposition 4, there will never be trade for prices exceeding  $\tilde{\omega}_1$  (because buyers will not accept such prices) and for prices lower than  $\tilde{\omega}_0$  (because sellers will not accept such prices). Combining the two results, we know that there exists an  $n_1$  such that for all  $n \geq n_1$ , the probability that any trade occurs at a price larger than the asset value in state  $Y$  (implying that some buyers – who endorse the correct state – make a loss) is at most  $\frac{\varepsilon}{2}$ , and analogously for state  $N$ .



**Event 3: no agent trades.** Let  $\mathcal{I} = [p^*, p^{**}]$ , let  $\mathcal{E}^+$  be the event that there is at least one buyer with  $p_i \in \mathcal{I}$  and  $\mathcal{E}^-$  be the event that there is at least one seller with  $p_i \in \mathcal{I}$  and  $\Delta_p \equiv p'' - p'$ . Then  $P(\mathcal{E}^+ | k \text{ buyers}) = 1 - (\Delta_p)^k$  and  $P(\mathcal{E}^- | n - k \text{ sellers}) = 1 - (\Delta_p)^{n-k}$ . Hence, under any state of the world  $S \in \{Y, N\}$ ,

$$\begin{aligned}
 & P(\text{trade in } \mathcal{I} | S) \\
 &= 1 - P(\text{no trade in } \mathcal{I} | S) \\
 &= 1 - \sum_{k=0}^n P(k \text{ buyers and } n - k \text{ sellers} | S) P(\text{not } \mathcal{E}^+ \text{ or not } \mathcal{E}^- | k \text{ buyers and } n - k \text{ sellers}) \\
 &\geq 1 - \sum_{k=0}^n \binom{n}{k} \omega_S^k (1 - \omega_S)^{n-k} [(\Delta_p)^k + (\Delta_p)^{n-k}] \\
 &= 1 - [\Delta_p \omega_S + (1 - \omega_S)]^n - [\omega_S + \Delta_p (1 - \omega_S)]^n.
 \end{aligned}$$

Since the limit of the last expression is 1,

$$\begin{aligned}
 & \lim_{n \rightarrow \infty} P(\text{no trade in } [p^*, p^{**}]) \\
 &= 1 - \lim_{n \rightarrow \infty} P(Y) P(\text{trade in } [p^*, p^{**}] | Y) + P(N) P(\text{trade in } [p^*, p^{**}] | N) \\
 &= 0.
 \end{aligned}$$

This implies that there exists an  $n_2 \geq n_\delta$  such that for all  $n \geq n_2$ , there is at most  $\frac{\varepsilon}{2}$  probability that no trade occurs in the interval  $(p^*, p^{**})$ .

$$\begin{aligned}
 & P(\text{trade in } \mathcal{I} | S) \\
 &\geq P(\mathcal{E}^+ \text{ and } \mathcal{E}^- | S) \\
 &= \sum_{k=1}^{n-1} P(k \text{ buyers and } n - k \text{ sellers} | S) P(\mathcal{E}^+ \text{ and } \mathcal{E}^- | k \text{ buyers and } n - k \text{ sellers}) \\
 &= \sum_{k=1}^{n-1} P(k \text{ buyers and } n - k \text{ sellers} | S) P(\mathcal{E}^+ | k \text{ buyers}) P(\mathcal{E}^- | k \text{ sellers}) \\
 &= 1 + (\Delta_p)^n - [\Delta_p \omega_S + (1 - \omega_S)]^{n-1} - [\omega_S + \Delta_p (1 - \omega_S)]^{n-1} \\
 &\quad - (1 - \omega_S)^n [1 - (\Delta_p)^k] [1 - (\Delta_p)^{n-k}].
 \end{aligned}$$

**Final step.** The probability that at least one agent has a positive payoff, and all those with positive payoffs have endorsed the actual state, and all those with negative payoffs have endorsed the opposite state, is at least  $1 - P(\text{Event 1}) - P(\text{Event 3})$ . For all  $n \geq n_1$ , the probability of Event 1 is at most  $\frac{\varepsilon}{2}$  and for all  $n \geq n_2$ , that of Event 3 is also at most  $\frac{\varepsilon}{2}$ . Let  $n^* = \max\{n_1, n_2\}$ . Then for any  $n \geq n^*$ , the probability that some agents endorsing the actual state, and only them, will make a profit and no agent endorsing the actual state makes a loss is more than  $1 - \varepsilon$ , as desired.

### 3.7.3 Parameter values for each task

In our current experimental design,  $P(S = Y) = P(S = N) = \frac{1}{2}$ , and group size  $n = 100$ . Table 3.7.1 lists parameter values  $\omega_N$  and  $\omega_Y$  of the 30 tasks in our experiment. They contain all combinations where  $(\omega_N, \omega_Y) \in \{0.05, 0.1, \dots, 0.45, 0.55, \dots, 0.95\}$ ,  $\omega_Y - \omega_N \in \{0.2, 0.25, 0.3\}$ , and  $\bar{\omega}_1 - \bar{\omega}_0 > 0.04$ .

Table 3.7.1: Task parameter values

set	task	$\omega_N$	$\omega_Y$
1	1	0.05	0.25
2	2	0.05	0.3
1	3	0.05	0.35
2	4	0.1	0.3
1	5	0.1	0.35
2	6	0.1	0.4
1	7	0.15	0.35
2	8	0.15	0.4
1	9	0.15	0.45
2	10	0.2	0.4
1	11	0.2	0.45
2	12	0.25	0.45
2	13	0.75	0.95
1	14	0.7	0.95
2	15	0.65	0.95
1	16	0.7	0.9
2	17	0.65	0.9
1	18	0.6	0.9
2	19	0.65	0.85
1	20	0.6	0.85
2	21	0.55	0.85
1	22	0.6	0.8
2	23	0.55	0.8
1	24	0.55	0.75
1	25	0.3	0.6
2	26	0.35	0.6
1	27	0.35	0.65
2	28	0.4	0.6
1	29	0.4	0.65
2	30	0.4	0.7

### 3.7.4 Truth-telling at the individual level

Table 3.7.2 shows the proportion of participants with at least certain numbers of truth-telling in both SPA and FTM treatments.

Table 3.7.2: Proportion of participants with at least certain numbers of truth-telling

at least	SPA	FTM
1	99.8%	100.0%
2	99.8%	99.8%
3	99.8%	99.8%
4	99.6%	99.6%
5	98.9%	98.9%
6	98.1%	98.3%
7	96.1%	96.6%
8	94.6%	94.1%
9	91.3%	91.3%
10	88.7%	85.8%
11	84.0%	82.2%
12	79.0%	77.6%
13	72.7%	71.9%
14	63.4%	64.3%
15	50.6%	53.1%

## 3.8 Proof of the lemmas

### 3.8.1 Proof of Lemma 1

*Proof.* Using that  $\tilde{P}_1(\omega_{-i} = 0) = \tilde{P}_1(\omega_{-i} = 1) = 0$ , since  $\tilde{P}$  conditions on the fact that  $\omega_{-i} \in (0, 1)$ , and further using that

$$\tilde{P}_1\left(\omega_{-i} = \frac{k}{n-1}\right) = \frac{P\left(\omega_{-i} = \frac{k}{n-1} \mid s_i = 1\right)}{1 - P(\omega_{-i} = 0 \text{ or } \omega_{-i} = 1 \mid s_i = 1)}$$

for  $0 < k < n - 1$ , we get

$$\begin{aligned}
 \tilde{\omega}_1 &= \sum_{k=1}^{n-2} \frac{k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \\
 &= \frac{1}{1 - P(\omega_{-i} = 0 \text{ or } \omega_{-i} = 1 \mid s_i = 1)} \sum_{k=1}^{n-2} \frac{k}{n-1} P \left( \omega_{-i} = \frac{k}{n-1} \mid s_i = 1 \right) \\
 &= \frac{1}{1 - P(\omega_{-i} = 0 \text{ or } \omega_{-i} = 1 \mid s_i = 1)} [E(\omega_{-i} \mid s_i = 1) - P(\omega_{-i} = 1 \mid s_i = 1)] \\
 &= \frac{1}{1 - P(\omega_{-i} = 0 \text{ or } \omega_{-i} = 1 \mid s_i = 1)} \left[ \frac{n}{n-1} \bar{\omega}_1 - \frac{1}{n-1} - P(\omega_{-i} = 1 \mid s_i = 1) \right].
 \end{aligned}$$

Note that in the last step, we use  $\omega_{-i} = \frac{n\omega-1}{n-1} = \frac{n}{n-1}\omega - \frac{1}{n-1}$  and Equation (3.2.3).

Since  $r_1, \omega_Y, \omega_N \in (0, 1)$ , as  $n \rightarrow \infty$ ,

$$\begin{aligned}
 &P(\omega_{-i} = 0 \mid s_i = 1) \\
 &= P(\omega_{-i} = 0 \mid Y)P(Y \mid s_i = 1) + P(\omega_{-i} = 0 \mid N)P(N \mid s_i = 1) \\
 &= r_1(1 - \omega_Y)^{n-1} + (1 - r_1)(1 - \omega_N)^{n-1} \rightarrow 0,
 \end{aligned}$$

and  $P(\omega_{-i} = 1 \mid s_i = 1) = r_1\omega_Y^{n-1} + (1 - r_1)\omega_N^{n-1} \rightarrow 0$ . We then have  $\tilde{\omega}_1 \rightarrow \bar{\omega}_1$ .  $\square$

### 3.8.2 Proof of Lemma 2

*Proof.* By the weak law of large numbers, we have that for any  $\delta > 0$  and  $\varepsilon > 0$  there exists  $n(\delta, \varepsilon)$  such that for all  $n \geq n(\delta, \varepsilon)$ :

$$P(|E[\omega_{-i} \mid Y] - \omega_{-i}| < \delta \mid Y) > 1 - \varepsilon. \quad (3.8.1)$$

Then note that  $E[\omega_{-i} \mid Y] = \omega_Y$ . For any  $\varepsilon > 0$ , taking  $\delta = \omega_Y - \bar{\omega}_1 > 0$  in Equation

(3.8.1), we have that for all  $n \geq n_1 \equiv n(\omega_Y - \bar{\omega}_1, \varepsilon)$ ,

$$\begin{aligned} 1 - \varepsilon &< P(|\omega_Y - \omega_{-i}| < \omega_Y - \bar{\omega}_1 | Y) \\ &\leq P(\omega_Y - \omega_{-i} < \omega_Y - \bar{\omega}_1 | Y) = P(\omega_{-i} > \bar{\omega}_1 | Y). \end{aligned}$$

Similarly, for all  $n \geq n_2 \equiv n(\bar{\omega}_0 - \omega_N, \varepsilon)$ ,  $P(\omega_{-i} < \bar{\omega}_0 | N) > 1 - \varepsilon$ .

From Lemma 1, for any  $\delta > 0$ , there exists  $m(\delta)$  such that  $|\tilde{\omega}_1 - \bar{\omega}_1| < \delta$  and  $|\tilde{\omega}_0 - \bar{\omega}_0| < \delta$  for all  $n \geq m(\delta)$ . Let us first pick  $\delta = \frac{\omega_Y - \bar{\omega}_1}{2} > 0$  and also in Equation (3.8.1). Then for all  $n \geq n_3 \equiv \max\{n(\frac{\omega_Y - \bar{\omega}_1}{2}, \varepsilon), m(\frac{\omega_Y - \bar{\omega}_1}{2})\}$ , we have

$$|\tilde{\omega}_1 - \bar{\omega}_1| < \frac{\omega_Y - \bar{\omega}_1}{2}$$

and

$$P\left(|\omega_Y - \omega_{-i}| < \frac{\omega_Y - \bar{\omega}_1}{2} | Y\right) > 1 - \varepsilon.$$

Hence,

$$\begin{aligned} P(\omega_{-i} > \tilde{\omega}_1 | Y) &\geq P(\omega_{-i} > \frac{\omega_Y + \bar{\omega}_1}{2} > \tilde{\omega}_1 | Y) \\ &= P(\omega_{-i} > \frac{\omega_Y + \bar{\omega}_1}{2} | Y) \\ &= P(\omega_Y - \omega_{-i} < \frac{\omega_Y - \bar{\omega}_1}{2} | Y) \\ &\geq P\left(|\omega_Y - \omega_{-i}| < \frac{\omega_Y - \bar{\omega}_1}{2} | Y\right) > 1 - \varepsilon. \end{aligned}$$

Similarly, for all  $n \geq n_4 \equiv \max\{n(\frac{\bar{\omega}_0 - \omega_N}{2}, \varepsilon), m(\frac{\bar{\omega}_0 - \omega_N}{2})\}$ ,  $P(\omega_{-i} < \tilde{\omega}_0 | N) > 1 - \varepsilon$ .

Finally, taking  $n^* = \max\{n_1, n_2, n_3, n_4\}$  completes the proof.  $\square$

### 3.8.3 Proof of Lemma 3

*Proof.* To prove the first part of the Lemma, first note that

$$\tilde{E}_s \left[ \omega_{-i} | \mathcal{E}_0^{-i}(p', p'') \right] = \sum_{k=1}^{n-2} \frac{k}{n-1} \tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} | \mathcal{E}_0^{-i}(p', p'') \right) \quad (3.8.2)$$

and, applying Bayes' rule,

$$\tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} | \mathcal{E}_0^{-i}(p', p'') \right) = \tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} \right) \frac{\tilde{P}_s \left( \mathcal{E}_0^{-i}(p', p'') | \omega_{-i} = \frac{k}{n-1} \right)}{\tilde{P}_s \left( \mathcal{E}_0^{-i}(p', p'') \right)}.$$

As we have

$$\begin{aligned} & \tilde{P}_s \left( \mathcal{E}_0^{-i}(p', p'') | \omega_{-i} = \frac{k}{n-1} \right) \\ &= \tilde{P}_s \left( \exists j \neq i, \text{ s.t. } s_j = 0 \text{ and } p' \leq p_j \leq p'' | \omega_{-i} = \frac{k}{n-1} \right) \\ &= 1 - \tilde{P}_s \left( \forall j \neq i \text{ s.t. } s_j = 0 : p_j < p' \text{ or } p_j > p'' | \omega_{-i} = \frac{k}{n-1} \right) \\ &= 1 - \left( 1 - (p'' - p') \right)^{n-k-1}, \end{aligned} \quad (3.8.3)$$

the expectation  $\tilde{P}_s \left[ \omega_{-i} | \mathcal{E}_0^{-i}(p', p'') \right]$  only depends on  $p'$  and  $p''$  through the difference  $\Delta_p \equiv p'' - p'$ . Consider next the likelihood ratio, for  $n-1 \geq k' > k \geq 2$ :

$$\frac{\tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} | \mathcal{E}_0^{-i}(p', p'') \right)}{\tilde{P}_s \left( \omega_{-i} = \frac{k'}{n-1} | \mathcal{E}_0^{-i}(p', p'') \right)} = \frac{1 - (1 - \Delta_p)^{n-k-1}}{1 - (1 - \Delta_p)^{n-k'-1}}$$

Taking derivative of this ratio with respect to  $\Delta_p$  gives

$$\frac{(1 - \Delta_p)^{n-k-2} \left[ 1 - (1 - \Delta_p)^{n-k'-1} \right] - (1 - \Delta_p)^{n-k'-2} \left[ 1 - (1 - \Delta_p)^{n-k-1} \right]}{\left[ 1 - (1 - \Delta_p)^{n-k'-1} \right]^2}$$

where the numerator can be simplified to

$$(1 - \Delta_p)^{n-k'-2} \left[ (1 - \Delta_p)^{k'-k} - 1 \right] < 0.$$

Thus, consider a point  $(q', q'')$  with  $\Delta_q \equiv q'' - q' > \Delta_p$ . Then,

$$\frac{\tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} \mid \mathcal{E}_0^{-i}(p', p'') \right)}{\tilde{P}_s \left( \omega_{-i} = \frac{k'}{n-1} \mid \mathcal{E}_0^{-i}(p', p'') \right)} > \frac{\tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} \mid \mathcal{E}_0^{-i}(q', q'') \right)}{\tilde{P}_s \left( \omega_{-i} = \frac{k'}{n-1} \mid \mathcal{E}_0^{-i}(q', q'') \right)}$$

which can be rewritten as

$$\frac{\tilde{P}_s \left( \omega_{-i} = \frac{k'}{n-1} \mid \mathcal{E}_0^{-i}(q', q'') \right)}{\tilde{P}_s \left( \omega_{-i} = \frac{k'}{n-1} \mid \mathcal{E}_0^{-i}(p', p'') \right)} > \frac{\tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} \mid \mathcal{E}_0^{-i}(q', q'') \right)}{\tilde{P}_s \left( \omega_{-i} = \frac{k}{n-1} \mid \mathcal{E}_0^{-i}(p', p'') \right)}$$

for all  $n-1 \geq k' > k \geq 2$ , whenever  $\Delta_q > \Delta_p$ . Hence, a monotone likelihood ratio property holds for the distributions  $\tilde{P}_s \left( \omega_{-i} \mid \mathcal{E}_0^{-i}(p', p'') \right)$  and  $\tilde{P}_s \left( \omega_{-i} \mid \mathcal{E}_0^{-i}(q', q'') \right)$ , implying that  $\tilde{E}_s \left[ \omega_{-i} \mid \mathcal{E}_0^{-i}(q', q'') \right] > \tilde{E}_s \left[ \omega_{-i} \mid \mathcal{E}_0^{-i}(p', p'') \right]$  if and only if  $\Delta_q > \Delta_p$ . As  $\Delta_p$  is increasing in  $p''$  and decreasing in  $p'$ , the first part of the lemma follows. The second part of the lemma can be obtained by replacing  $n-k-1$  by  $k$  in the derivations above, changing the signs of all derivatives.  $\square$

### 3.8.4 Proof of Lemma 4

*Proof.* Recall that  $\tilde{P}_1(\cdot)$  conditions on  $k \leq n-2$  and thus  $\tilde{P}_1(\exists j \neq i : s_j = 0) = 1$ . Hence,  $\tilde{P}_1 \left( \mathcal{E}_0^{-i}(0, 1) \mid \omega_{-i} = \frac{k}{n-1} \right) = 1$  for all  $k \leq n-2$  and  $\tilde{P}_1 \left( \mathcal{E}_0^{-i}(0, 1) \right) = 1$ . Hence,

$$\tilde{E}_1 \left[ \omega_{-i} \mid \mathcal{E}_0^{-i}(0, 1) \right] = \sum_{k=1}^{n-2} \frac{k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) = \tilde{\omega}_1.$$

By the same reasoning,  $\tilde{\omega}_1$  is also equal to  $\tilde{E}_1 \left[ \omega_{-i} \mid \mathcal{E}_1^{-i}(0, 1) \right]$ . As from Lemma 3,

$$\tilde{E}_1 \left[ \omega_{-i} \mid \mathcal{E}_0^{-i}(0, 1) \right] > \tilde{E}_1 \left[ \omega_{-i} \mid \mathcal{E}_0^{-i}(p', p'') \right]$$



and

$$\tilde{E}_1 \left[ \omega_{-i} | \mathcal{E}_1^{-i}(0, 1) \right] < \tilde{E}_1 \left[ \omega_{-i} | \mathcal{E}_1^{-i}(p', p'') \right]$$

whenever  $p'' - p' < 1$ , the Lemma follows.  $\square$

### 3.8.5 Proof of Lemma 5

*Proof.* Consider Equation (3.8.2). Only the weights  $\frac{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p'') | \omega_{-i} = \frac{k}{n-1})}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p''))}$  are functions of  $p'$  and  $p''$ . Since the denominator is the weighted sum of the numerator for all  $k$ , it is enough to show that  $\tilde{P}_1(\mathcal{E}_0^{-i}(p', p'') | \omega_{-i} = \frac{k}{n-1})$  is continuous in  $p'$  and  $p''$ . This is immediate from Equation (3.8.3).  $\square$

### 3.8.6 Proof of Lemma 6

*Proof.* We start from Equation (3.8.2) and first show that the denominator of the fraction in the expression goes to 1 as  $n \rightarrow \infty$ . Using Equation (3.8.3) and since for any integer  $k \in [1, n-2]$ ,

$$\begin{aligned} & \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \\ &= \binom{n-1}{k} \omega_Y^k (1 - \omega_Y)^{n-k-1} \tilde{r}_1 + \binom{n-1}{k} \omega_N^k (1 - \omega_N)^{n-k-1} (1 - \tilde{r}_1), \end{aligned} \tag{3.8.4}$$

we arrive at

$$\begin{aligned}
 & \tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \right) \\
 &= \sum_{k=1}^{n-2} \tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \mid \omega_{-i} = \frac{k}{n-1} \right) \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \\
 &= \sum_{k=1}^{n-2} \left[ 1 - (1 - \Delta_p)^{n-k-1} \right] \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \\
 &= 1 - \sum_{k=1}^{n-2} (1 - \Delta_p)^{n-k-1} \binom{n-1}{k} \left[ \omega_Y^k (1 - \omega_Y)^{n-k-1} \tilde{r}_1 + \omega_N^k (1 - \omega_N)^{n-k-1} (1 - \tilde{r}_1) \right] \\
 &= 1 - \tilde{r}_1 \left( [\omega_Y + (1 - \Delta_p)(1 - \omega_Y)]^{n-1} - \omega_Y^{n-1} - (1 - \Delta_p)^{n-1} (1 - \omega_Y)^{n-1} \right) \\
 &\quad - (1 - \tilde{r}_1) \left( [\omega_N + (1 - \Delta_p)(1 - \omega_N)]^{n-1} - \omega_N^{n-1} - (1 - \Delta_p)^{n-1} (1 - \omega_N)^{n-1} \right),
 \end{aligned} \tag{3.8.5}$$

where the last equality is obtained from the binomial theorem. In the limit, for any prices  $p', p''$  in  $(0, 1)$ ,

$$\lim_{n \rightarrow \infty} \tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \right) = 1 \tag{3.8.6}$$

since  $p', p'', \omega_Y$ , and  $\omega_N$  are all strictly between 0 and 1.

Next, we bound the expression in Equation (3.8.2) from above and below to show that it tends to  $\tilde{\omega}_1$  in the limit. Lemma 4 already gives the upper bound,  $\tilde{\omega}_1$ . For the lower bound:

$$\begin{aligned}
 & \tilde{E}_1 \left[ \omega_{-i} \mid \mathcal{E}_0^{-i} (p', p'') \right] \\
 &= 1 - \tilde{E}_1 \left[ 1 - \omega_{-i} \mid \mathcal{E}_0^{-i} (p', p'') \right] \\
 &= 1 - \frac{\sum_{k=1}^{n-2} \left( 1 - \frac{k}{n-1} \right) \tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \mid \omega_{-i} = \frac{k}{n-1} \right) \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right)}{\tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \right)} \\
 &> 1 - \frac{\sum_{k=1}^{n-2} \left( 1 - \frac{k}{n-1} \right) \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right)}{\tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \right)} \\
 &= 1 - \frac{1 - \tilde{\omega}_1}{\tilde{P}_1 \left( \mathcal{E}_0^{-i} (p', p'') \right)}.
 \end{aligned} \tag{3.8.7}$$

Combining the two bounds, we get

$$1 - \frac{1 - \tilde{\omega}_1}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p''))} < \tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] < \tilde{\omega}_1. \quad (3.8.8)$$

Taking the limit and using Equation (3.8.6) and Lemma 1, we get that for any  $p', p''$  in  $(0, 1)$ ,  $\lim_{n \rightarrow \infty} \tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] = \lim_{n \rightarrow \infty} \tilde{\omega}_1 = \bar{\omega}_1$ .  $\square$

### 3.8.7 Proof of Lemma 7

*Proof.* From Lemma 6, we know that

$$\lim_{n \rightarrow \infty} \tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] = \bar{\omega}_1$$

and

$$\lim_{n \rightarrow \infty} \tilde{E}_0[\omega_{-i} | \mathcal{E}_1^{-i}(p', p'')] = \bar{\omega}_0.$$

We define a function  $g$  mapping all prices  $p', p''$  such that  $p'' > p'$  into  $[-1, 1]^2$ , with

$$g(p', p'') = \left( \tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(p', p'')] - p'', p' - \tilde{E}_0[\omega_{-i} | \mathcal{E}_1^{-i}(\alpha, 1)] \right).$$

It immediately follows from Lemma 5 that  $g$  is continuous. Since we also have  $0 < \bar{\omega}_0 < \bar{\omega}_1 < 1$ , we get that there must be  $\alpha \in (\bar{\omega}_0, \bar{\omega}_1)$  such that  $\bar{\omega}_0 < \alpha < \bar{\omega}_1$  for sufficiently large  $n$ ,

$$\tilde{E}_1[\omega_{-i} | \mathcal{E}_0^{-i}(\alpha, 1)] - 1 < 0,$$

$$\alpha - \tilde{E}_0[\omega_{-i} | \mathcal{E}_1^{-i}(\alpha, 1)] < 0,$$

and

$$\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(0, \alpha)] - \alpha > 0,$$

$$0 - \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(0, \alpha)] > 0.$$

By the intermediate value theorem, there must then exist a pair  $p^*, p^{**}$  with  $0 < p^* < \alpha < p^{**} < 1$  with the property:

$$\tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p^*, p^{**})] - p^{**} = 0$$

$$p^* - \tilde{E}_0 [\omega_{-i} | \mathcal{E}_1^{-i}(p^*, p^{**})] = 0.$$

□

### 3.8.8 Proof of Lemma 8

*Proof.* Consider the following expectation

$$\begin{aligned} \tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p', p)] &= \sum_{k=1}^{n-2} \frac{k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} | \mathcal{E}_0^{-i}(p', p) \right) \\ &= \sum_{k=1}^{n-2} \frac{k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \frac{\tilde{P}_1 \left( \mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{k}{n-1} \right)}{\tilde{P}_1 \left( \mathcal{E}_0^{-i}(p', p) \right)}. \end{aligned}$$

Taking derivative with respect to  $p$  gives:

$$\begin{aligned} &\frac{\partial}{\partial p} \tilde{E}_1 [\omega_{-i} | \mathcal{E}_0^{-i}(p', p)] \\ &= \sum_{k=1}^{n-2} \frac{k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \frac{\partial}{\partial p} \left[ \frac{\tilde{P}_1 \left( \mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{k}{n-1} \right)}{\tilde{P}_1 \left( \mathcal{E}_0^{-i}(p', p) \right)} \right] \\ &= \sum_{k=1}^{n-2} \frac{n-1-k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{n-1-k}{n-1} \right) \frac{\partial}{\partial p} \left[ \frac{\tilde{P}_1 \left( \mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1} \right)}{\tilde{P}_1 \left( \mathcal{E}_0^{-i}(p', p) \right)} \right]. \end{aligned} \tag{3.8.9}$$

where the second step exchanges the order of summation. Consider next exclusively the term  $\frac{\partial}{\partial p} \left[ \frac{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1})}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} \right]$ . Expanding yields:

$$\frac{\frac{\partial}{\partial p} \tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1}) \tilde{P}_1(\mathcal{E}_0^{-i}(p', p)) - \tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1}) \frac{\partial}{\partial p} \tilde{P}_1(\mathcal{E}_0^{-i}(p', p))}{[\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))]^2}.$$

Since  $\frac{\partial}{\partial p} \tilde{P}_1(\mathcal{E}_0^{-i}(p', p))$  is strictly positive, this is strictly smaller than

$$\begin{aligned} & \frac{\frac{\partial}{\partial p} \tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1}) \tilde{P}_1(\mathcal{E}_0^{-i}(p', p))}{[\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))]^2} \\ &= \frac{\frac{\partial}{\partial p} \tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1})}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} \end{aligned}$$

and since

$$\tilde{P}_1\left(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1}\right) = 1 - \left(1 - (p - p')\right)^{n-1-(n-1-k)} = 1 - \left(1 - (p - p')\right)^k,$$

this last expression equals:

$$\frac{k(1 - (p - p'))^{k-1}}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))}. \quad (3.8.10)$$

Note that the expression in 3.8.10 converges to 0 as  $k \rightarrow \infty$ . For any given  $n$  (which fixes  $\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))$ ), there thus has to exist  $k^*$  such that

$$\frac{k(1 - (p - p'))^{k-1}}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} < 1 \quad (3.8.11)$$

for all  $k \geq k^*$ . Note further that since the numerator is fixed in  $n$  and  $\tilde{P}_1(\mathcal{E}_0^{-i}(p^*, p))$  is increasing in  $n$ , this  $k^*$  will have the same property for all  $n' \geq n$ . We can next split the

sum in Equation (3.8.9) into two parts:

$$\begin{aligned}
 & \sum_{k=1}^{k^*} \frac{n-1-k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{n-1-k}{n-1} \right) \frac{\partial}{\partial p} \left[ \frac{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1})}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} \right] \\
 + & \sum_{k=k^*+1}^{n-1} \frac{n-1-k}{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{n-1-k}{n-1} \right) \frac{\partial}{\partial p} \left[ \frac{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1})}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} \right] \\
 < & \sum_{k=1}^{k^*} \tilde{P}_1 \left( \omega_{-i} = \frac{k}{n-1} \right) \frac{k(1-(p-p'))^{k-1}}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} \\
 + & \sum_{k=k^*+1}^{n-1} \tilde{P}_1 \left( \omega_{-i} = \frac{n-1-k}{n-1} \right) \frac{\partial}{\partial p} \left[ \frac{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p) | \omega_{-i} = \frac{n-1-k}{n-1})}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} \right] \\
 < & \tilde{P}_1 \left( \omega_{-i} \leq \frac{k^*}{n-1} \right) \max_{\ell} \frac{\ell(1-(p-p'))^{\ell-1}}{\tilde{P}_1(\mathcal{E}_0^{-i}(p', p))} + (1 - \varepsilon_n)
 \end{aligned}$$

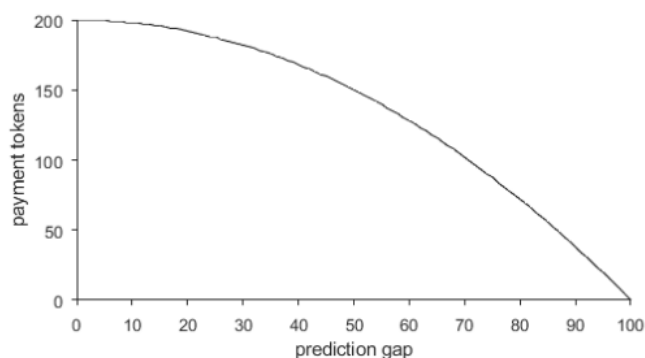
where  $\varepsilon_n > 0$  since by construction, the second sum averages over values smaller than 1. Since we also have  $\tilde{P}_1 \left( \omega_{-i} \leq \frac{k^*}{n-1} \right)$  converging to 0 as  $n \rightarrow \infty$ , there exists  $n^*$  such that for all  $n \geq n^*$ , the first part of the sum is smaller than  $\varepsilon_n$ , and thus the lemma follows.  $\square$

### 3.9 Experimental instructions and training rounds

All participants first watched the experimental instruction video (YouTube link) where the experimental setting of urns and balls was explained.

Then they went through five rounds of training and got feedback about how the payment was calculated. Figure 3.9.1 shows an example from the SPA treatment, and Figure 3.9.2 shows an example from the FTM treatment where the bet went through.

Your payment for this task depends on your prediction gap, which is the distance between your prediction and the actual number of participants who have guessed the same urn as you. **The less the prediction gap is, the higher your payment will be.**



**Your payment for this task: 182 tokens**

At the end of the experiment, we sum up your total payment tokens from all the 15 tasks and determine your bonus. Your bonus (in £) equals the total number of payment tokens divided by 1000. You could earn up to £3 in bonus.

Figure 3.9.1: An example of feedback in the training rounds in SPA.

The earnings of the bet (in tokens) is this number minus 55:

$$\text{Earnings} = 80 - 55 = 25$$

To make sure you don't lose money (if the earnings are negative), you will be endowed with 100 tokens, whether you take the bet or not.

You chose to take the bet.

Your bet goes through if someone took the opposite bet, which means, someone bet that less than 55 participants chose Urn Right (or in other words, more than 45 chose the other urn).

**Your bet went through. Your payment for this task is  $100 + \text{earnings} = 125$  tokens.**

If you had not taken the bet, then your payment would have been the endowment, 100 tokens.

At the end of the experiment, we sum up your total payment tokens from all the 15 tasks and determine your bonus. Your bonus (in £) equals the total number of payment tokens divided by 1000. You could earn up to £3 in bonus.

**Note that the value 55 in the bet is randomly drawn.**

Figure 3.9.2: An example of feedback in the training rounds in FTM.



# Chapter 4

## Honesty via Choice-Matching<sup>1</sup>

### 4.1 Introduction

More than other social sciences, economics favors the hard data revealed by agents' decisions, as opposed to surveys of their opinions or tastes. Yet subjective reports are often an irreplaceable source of information. Firms that want to know how consumers would value a hypothetical new product, government agencies asking external experts to assess the long-term impact of a policy, and economists working with self-reported panel-data all rely on answers provided by individuals whose effort or honesty may be in doubt. We propose a new class of incentive compatible mechanisms for this problem that we call *choice-matching*. The idea is to link explicit opinions and judgments, with an auxiliary task that reveals the respondents' "types", but only implicitly. In our canonical version, respondents answer a multiple choice question (MCQ), and the auxiliary task asks respondents to predict how often each answer was chosen by all other respondents.

A respondent's score is then a weighted sum of a prediction accuracy score and the average prediction accuracy score of all the respondents who endorse the same answer to the MCQ. We state conditions under which this mechanism is truth-inducing, the major

---

<sup>1</sup>This chapter is based on an eponymous paper co-authored with Jakša Cvitanić, Dražen Prelec and Blake Riley.

one being that there is a one-to-one correspondence between predictions and answers to the MCQ. However, the guiding idea behind choice-matching is more general, and the prediction question can be replaced with other kinds of tasks or games.

We illustrate choice-matching with an example of consumer evaluations of a trial product in Section 4.2. Section 4.3 presents the model and our results in the canonical version of our mechanism. Section 4.4 explains how the choice-matching payment rules can be used without asking for predictions of other respondents' answers, and discusses possible applications in marketing and experimental economics. Section 4.5 compares our method to previous proposals. Section 4.6 concludes. Additional material is provided in Section 4.7

## 4.2 Example

Suppose that a firm intends to launch a new product and asks respondents to evaluate a trial version on a scale from 1 to 5. A rating that best represents a respondent's true opinion is their *type*. The firm would like to know the actual percentage of each type, but may be worried that some respondents may not answer honestly, for lack of effort or because they feel obliged to endorse the product. The firm wants to provide monetary incentives for honest answers. For these incentives to be strict, there must be some additional input to the mechanism beyond the answers themselves (Radanovic and Faltings, 2013; Cvitanović et al., 2020). In our canonical version of choice-matching, this input is a respondent's prediction of the relative frequency of each of the five ratings submitted by other respondents in the sample.

Let  $x^r = (x_1^r, \dots, x_5^r)$  denote the reported rating of respondent  $r$ , where  $x_k^r = 1$  if  $r$ 's answer is " $k$ " and  $x_k^r = 0$  otherwise.  $y^r = (y_1^r, \dots, y_5^r)$ , is then a vector of predictions, where  $y_k^r$  is the prediction by respondent  $r$  of the frequency of others choosing  $k$  as their MCQ answer.

The “prediction score” that respondent  $r$  receives is a function  $S(\bar{x}^{-r}, y^r)$  of her prediction vector  $y^r$  and the distribution  $\bar{x}^{-r}$  of the reported ratings’ frequencies, excluding her own reported rating. The firm chooses  $S$  to be a proper scoring rule (Savage, 1971, Gneiting and Raftery, 2007). It is then in the interest of each respondent to report their expected value of the frequencies, which we call their *best prediction*. We denote by  $\bar{S}^{-r}$  the average prediction score of respondents other than  $r$  who report the same rating as  $r$ , if any. Respondent  $r$  receives a score of zero if among the others’ ratings there is any missing ratings option, not reported by at least one respondent other than  $r$ . If there is no missing option, she receives

$$\lambda S + (1 - \lambda) \bar{S}^{-r}$$

with  $\lambda \in (0, 1)$  an arbitrary weight chosen by the survey planner.

Why might we expect this payment formula to be truth-inducing? Observe, first, that respondent  $r$  can safely assume that the payment formula is in effect, because the presence or absence of a missing option does not hinge on her rating, and if the formula is not in effect she receives zero no matter what she does. This triggering mechanism was first employed by Baillon (2017), for the binary case.

The term  $\lambda S$  provides incentives for giving best predictions because  $S$  is a strictly proper scoring rule. If a respondent believes that those who share her type also have (roughly) the same expectations of type frequencies, she will anticipate that they will also give (roughly) the same predictions. Then, if all respondents other than her provide honest ratings, the average of the prediction scores of all those with the same rating is (roughly) equal to the score corresponding to their best prediction. By providing her honest rating, a respondent is scored according to a prediction (roughly) identical to the one she made herself, ensuring that the second term provides the incentives to answer the MCQ honestly. These incentives apply even if some respondents do not provide a personal prediction. Thus, in a large survey most respondents could be asked just for their ratings

(letting  $\lambda = 0$  in their scoring formula).

For the above arguments, it is not crucial that respondents use what we call best predictions. What is important is that respondents of the same type believe that approximately the same predictions are optimal, and respondents of different types believe that different predictions are optimal, as we discuss below. Previous mechanisms have also invoked this assumption in some form, but with additional requirements or limitations (Prelec, 2004; Witkowski and Parkes, 2012b; Radanovic and Faltings, 2013; 2014; Zhang and Cheng, 2014; Baillon, 2017). We will compare choice-matching to these proposals at the end of the chapter.

## 4.3 Model and Results

### 4.3.1 Setup

Generalizing the previous example, let  $A = \{1, \dots, M\}$  denote the set of possible answers to an MCQ, and let  $N$  be the number of respondents. The respondent's *type* is denoted by random vector  $T^r$ , where  $T^r = i$  if  $i \in A$  would be respondent  $r$ 's honest answer to the question, and  $T_i^r = 0$  otherwise.<sup>2</sup>  $X^r = (X_1^r, \dots, X_M^r)$  denotes the chosen answer or *reported type* of respondent  $r$ , where  $X_k^r = 1$  if  $r$ 's answer is " $k$ " and  $X_k^r = 0$  otherwise. The answer is *honest* iff  $X^r = T^r$ . For  $i \in A$ ,  $\bar{T}_i^{-r} = \frac{1}{N-1} \sum_{s' \neq r} T_i^{s'}$  is the frequency of types  $i$  in the sample, after excluding respondent  $r$ , and similarly,  $\bar{X}_i^{-r} = \frac{1}{N-1} \sum_{s \neq r} X_i^s$  is the frequency of reported types  $i$ .  $Y_i^r$  is respondent  $r$ 's prediction of  $\bar{X}_i^{-r}$ . The realizations of  $X_i^r$  and  $Y_i^r$  are denoted  $x_i^r, y_i^r$ , the realization of  $T_i^r$  is  $t_i^r$ , and so on.

A respondent's score will depend on the reported types and predictions of all respondents, and the following choice-matching trigger:

**Definition 18.** The *choice-matching trigger*  $\mathcal{M}^r$  of respondent  $r$  is the event such that

---

<sup>2</sup>We will say in this case that  $r$  has type  $i$ .

each answer  $i \in A$  is reported at least once by respondents other than  $r$ . That is,  $\mathcal{M}^r$  occurs if and only if there is no  $i \in A$  with  $\bar{x}_i^{-r} = 0$ . The *type-matching trigger*  $\mathcal{E}^r$  of respondent  $r$  is the event such that each type  $i \in A$  is represented at least once among respondents other than  $r$ .

Note that when all respondents other than  $r$  respond honestly, the choice-matching trigger becomes identical to the type-matching trigger. Assumptions A4-A7 below are sufficient, but not necessary, for choice-matching to be honesty-inducing. In Subsection 4.3.2 we explain how they can be relaxed.

**Assumption 4. *Common prior.*** *There exists a common prior on the distribution of  $T^r$  in the population.*

Importantly, we do not assume that the survey planner knows the prior, only that it is known to the respondents.

**Assumption 5. *Non-degeneracy.*** *For any respondent  $r$  and any realization  $t^r$ :*

$$P(\mathcal{E}^r \mid T^r = t^r) > 0$$

Each respondent considers it possible that all distinct types are represented in the rest of the sample. Although a technical assumption, non-degeneracy cannot be relaxed without changing our approach substantially. The method thus puts some constraints on the survey design: the number of answer options must be less than the number of respondents (continuous scales are not allowed), and all options must have a chance of receiving an honest endorsement.

We now write  $r$ 's expectation of others' type frequencies, conditioning on own type  $k$  and  $\mathcal{E}^r$ , as

$$p^{r,k} = E[\bar{T}^{-r} \mid T_k^r = 1, \mathcal{E}^r]$$

**Assumption 6. *Stochastic relevance.*** For any two respondents  $r, s$  and any answer options,  $k, i \in A$ :

$$p^{r,k} \neq p^{s,i} \text{ if } k \neq i$$

**Assumption 7. *Impersonal updating.*** For any two respondents  $r, s$  and any answer option,  $k \in A$ :

$$p^{r,k} = p^{s,k}$$

Our version of stochastic relevance differs slightly from the standard one (Miller et al., 2005), because expectations are also conditioned on the type-matching trigger  $\mathcal{E}^r$  (this further implies  $N \geq M + 2$ , since  $N = M + 1$  and  $\mathcal{E}^r$  yield  $p^{r,k} = (1/M, \dots, 1/M)$ , for all  $r, k$ ).

Stochastic relevance is a mild requirement, with substantial support in the experimental psychology literature (Marks and Miller, 1987). The finding there that respondents' answers about their personal characteristics are strongly correlated with their predictions about the distribution of answers in the sample was labeled the “false consensus effect” (Ross et al., 1977), although, as noted by Dawes (1989), it can be consistent with Bayesian updating.

Stochastic relevance may be violated if the distribution of a particular characteristic – for instance gender – is common knowledge. However, such a characteristic can be made stochastically relevant if combined with characteristics whose distribution is not common knowledge. For example, a respondent could be asked to indicate both gender and beverage preference, and to predict their joint distribution in the sample. If the correlation between beverage preference and gender is not known a priori, then one's own gender will be informative of the joint gender-preference distribution.

Impersonal updating is a more demanding assumption, stating that all respondents of the same type have identical expectations about type frequencies. It will not hold exactly in practice, as respondents with identical answers to an MCQ are likely to report differing

predictions. As we discuss below, the main result of this section continues to hold when impersonal updating is only approximately true.

Assumptions A4-A7 generically hold in a setting in which respondents' types are independently and identically distributed conditional on a certain state of the world.<sup>3</sup> In our example in 4.2 for instance, we could imagine that different states of the world correspond to the true quality of the trial product and that the better the quality, the higher the probability for each respondent to have a high honest rating.<sup>4</sup>

### 4.3.2 Inducing Honesty via Choice-Matching

We now assume A4-A7 and model the strategic setting induced by our payment rule as a Bayesian game. A pure strategy for respondent  $r$  is a function  $\sigma(t^r) = (\sigma_x(t^r), \sigma_y(t^r))$  that maps her type to a response  $(x^r, y^r)$ . The profile of all respondents' pure strategies is denoted  $\sigma(t)$ , with entries  $\sigma^r(t^r)$ , and the profile excluding player  $r$  is denoted  $\sigma^{-r}(t^{-r})$ .

Given a real-valued payment rule  $R(\sigma^r(t^r), \sigma^{-r}(t^{-r}))$ , a set of response strategies is a (Bayesian) *Nash equilibrium*, if, for any response  $(x, y) \neq (\sigma^r(t^r))$ , we have

$$E \left[ R \left( \sigma_x^r(t^r), \sigma_y^r(t^r); \sigma^{-r}(t^{-r}) \right) - R \left( x, y; \sigma^{-r}(t^{-r}) \right) \mid T^r = t^r \right] \geq 0 \quad (4.3.1)$$

That is, by deviating in responses  $(x, y)$ , player  $r$  would be worse off (in expectation) than by not deviating. If the inequality is strict we speak of a *strict* Nash equilibrium. A Nash equilibrium is strict in  $x$  if the inequality is strict whenever  $x \neq \sigma_x^r(t^r)$  and, analogously, it is strict in  $y$  if the inequality is strict whenever  $y \neq \sigma_y^r(t^r)$ .

---

<sup>3</sup>An exact statement on the relation of the conditional i.i.d. assumption and our assumptions is given in Section 4.7. One sufficient condition for our assumptions to hold then is that  $N > M + 1$  and that the state of the world has a continuous distribution.

<sup>4</sup>Furthermore, our assumptions can hold even if the exact distribution of types is common knowledge. For example, with  $N = 4$ ,  $M = 2$  and common knowledge that there are two respondents of each type, each type knows that among the remaining three respondents there is one that matches her type and two who do not. A pair of respondents will therefore have the same expectations if and only if the pair are of the same type, satisfying assumptions A6 and A7.

A strategy profile is *honest* if every respondent  $r$  reports  $x^r = t^r$ . A payment rule which has a Nash equilibrium that is strict in  $x$  and  $y$  and honest is called *strictly incentive compatible*. To make choice-matching incentive compatible, we use strictly proper scoring rules:

**Definition 19.** Let  $Z$  be a random vector of positive frequencies with dimension  $M > 1$  which satisfies  $\sum_{k=1}^M Z_k = 1$ . Let  $y$  be a prediction of  $Z$ . We say that the scoring rule  $S$  is *strictly proper* if  $E[S(y, Z)]$  is uniquely maximized for  $y = E[Z]$ .

Well-known strictly proper scoring rules are the quadratic  $S(y, Z) = -\sum_{k=1}^M (y_k - Z_k)^2$ , and the logarithmic  $S(y, Z) = \sum_{k=1}^M Z_k \log(y_k)$ .

**Definition 20.** Consider a respondent  $r$  who reports  $k$  as her type, and let  $S$  be strictly proper. We say that a collection of payment rules  $R_{S,\lambda}^r$  induces *choice-matching* if

(a) In the event  $\mathcal{M}^r$ :

$$R_{S,\lambda}^r(x^r, y^r, x^{-r}, y^{-r}) = \lambda S(y^r, \bar{x}^{-r}) + (1 - \lambda) \bar{S}^{-r}(x^r, x^{-r}, y^{-r})$$

where  $\lambda \in (0, 1)$  and  $\bar{S}^{-r}(x^r, x^{-r}, y^{-r})$  is the average prediction score achieved by the respondents other than  $r$  who submit  $x^s = x^r$ :

$$\bar{S}^{-r}(x^r, x^{-r}, y^{-r}) = \frac{\sum_{s \neq r} x^r \cdot x^s S(y^s, \bar{x}^{-s})}{\sum_{s \neq r} x^r \cdot x^s}$$

(b) and  $R_{S,\lambda}^r(x^r, y^r) = 0$  otherwise.

In words, if all  $M$  possible answers are chosen by at least one respondent other than  $r$ , choice-matching assigns her a score that is a weighted average of her own prediction score (that is,  $S(y^r, \bar{x}^{-r})$ ) and the prediction score of those respondents who report the same type. Otherwise, she receives zero.

We now come to the main result of this section.



**Proposition 5.** *Under Assumptions A4-A7 any collection of payment rules  $R_{S,\lambda}^r$  that induces choice-matching is strictly incentive compatible.*

*Proof.* We show that providing honest responses and best predictions is a Nash equilibrium that is honest and strict in  $x$  and  $y$ , hence strictly incentive compatible. Fix a respondent  $r$  with type  $k$  and suppose that all respondents other than  $r$  provide honest responses and best predictions. Because respondent  $r$  cannot influence the choice-matching trigger and receives zero if she is not matched, she will condition her payoffs on being choice-matched. Given that other respondents answer honestly, this means conditioning on the type-matching event  $\mathcal{E}^r$ . With her prediction  $y^r$ ,  $r$  thus maximizes

$$E[S(y^r, \bar{T}^{-r}) | \mathcal{E}^r, T_k^r = 1]$$

Since  $S$  is strictly proper,  $r$  maximizes her expected payoff with  $y^r = p^{r,k}$ , i.e., what we call  $r$ 's "best prediction" is the optimal prediction from  $r$ 's perspective.<sup>5</sup> To see that choice-matching is strictly incentive compatible in  $x$ , consider the difference in the expected score for respondent  $r$  between reporting  $t^r$  honestly and deviating by making some dishonest report  $x^r$  with  $x_i^r = 1$ :

$$Pr(\mathcal{E}^r | T^r) \times (1 - \lambda) E[\bar{S}^{-r}(t^r, t^{-r}, y^{-r}) - \bar{S}^{-r}(x^r, t^{-r}, y^{-r}) | T^r = t^r, \mathcal{E}^r]$$

From non-degeneracy,  $Pr(\mathcal{E}^r | T^r) \times (1 - \lambda) > 0$ . Furthermore, by impersonal updating all respondents of the same type have the same best prediction, so that the average  $\bar{S}^{-r}$  is equal to the corresponding score  $S$ , and we have

$$E[\bar{S}^{-r}(t^r, t^{-r}, y^{-r}) - \bar{S}^{-r}(x^r, t^{-r}, y^{-r}) | T^r = t^r, \mathcal{E}^r] = E[S(p^{r,k}, \bar{x}^{-r}) - S(p^{r,i}, \bar{x}^{-r}) | T^r = t^r, \mathcal{E}^r]$$

---

<sup>5</sup>Due to non-degeneracy,  $\tilde{E}[\cdot] := E[\cdot | \mathcal{E}^r, T_k^r = 1]$  is also an expectation operator, so that when  $S$  is strictly proper, for any random vector of frequencies  $Z$ ,  $\tilde{E}[S(r, Z)]$  is maximized by  $r = \tilde{E}[Z]$ .

By definition of  $p^{r,k}$  and since  $S$  is strictly proper, we have

$$E[S(p^{r,k}, \bar{x}^{-r}) - S(p^{r,i}, \bar{x}^{-r}) \mid T^r = t^r, \mathcal{E}^r] > 0$$

for any  $i \neq k$ . Thus,  $R_{S,\lambda}$  is strictly incentive compatible.  $\square$

### 4.3.3 Robustness of the Result

In the following, we discuss extensions of Proposition 5. First, there may be equilibria other than the honest one. However, if respondents cannot communicate with each other, it is plausible that respondents of the same type will adopt the same strategies (perhaps mixed). Under this *symmetry* assumption, we can show that all strict equilibria are either honest, or honest up to permutation of responses, that is, all respondents of type  $k$  report type  $i$ , all respondents of type  $i$  report type  $j$ , and so on. Second, we discuss in which way assumptions A4-A7 can be relaxed, while still allowing for a strict Nash equilibrium in honest strategies. Third, we point out that it is not necessary that all respondents report both an answer to the MCQ and a prediction, and that proper scoring can be replaced with other belief elicitation mechanisms.

A strategy profile  $\sigma$  is *symmetric* if respondents of the same type have identical strategies, i.e., if for all respondents  $r$  and  $s$  and all types  $t^r, t^s$  we have  $\sigma^r(t^r) = \sigma^s(t^s)$  whenever  $t^s = t^r$ . The profile  $\tilde{\sigma}$  is a *permutation profile* of  $\sigma$  if for any respondent  $r$  and any  $k \in A$ , we have  $\tilde{\sigma}_{\pi(k)}(t^t) = \sigma_k(t^r)$ , where  $\pi$  is a permutation of the answers to the MCQ.

**Corollary 1.** *Any permutation profile of honest responses is a strict Nash equilibrium. Moreover, when  $\sigma$  is a symmetric strict Nash equilibrium, then  $\sigma$  is a permutation profile of honest responses.*

*Proof.* The first part of Corollary 1 follows directly from the proof of Proposition 5, replacing  $k$  by  $\pi(k)$ . The second part rules out pooling equilibria as follows. Since  $\sigma$

is symmetric, respondents of the same type always provide the same answers. First, if respondents of different types report different answers, the reported types  $\sigma_x(t)$  are a permutation of the (actual) type profile. As in the proof of Proposition 5, any respondent  $r$  of type  $k$  then maximizes her expected prediction score by setting  $y_\ell^r = p_{\sigma_x^{-1}(\ell)}^{r,k}$  for all  $\ell \in A$ . Taken together,  $\sigma(t)$  is a permutation profile of honest responses. Second, if respondents of two different types report the same type, there must be a missing answer option and thus all respondents receive zero. Since respondents could also receive zero by endorsing the missing answer option, the equilibrium cannot be strict.  $\square$

Permutation profiles are a relabeling of the answers to the MCQ, and provide no benefit in score relative to the focal honest equilibrium. It is hard to see how such relabeling might be individually or collectively advantageous. Pooling equilibria may be attractive to respondents who would like to conceal their type. However, as the above proof shows, they are either not strict or not symmetric (or both).

Proposition 5 and Corollary 1 still hold when we weaken the impersonal updating assumption A7. Note first that A7 does not affect incentives for the predictions. If all respondents other than  $r$  are honest,  $r$  should report her best prediction  $p^{r,k}$  when her type is  $k$ . It is intuitive that  $r$  will have strict incentives to truthfully report her type if the best prediction of all individuals of her type is “closer” to her best prediction than the best prediction of individuals of some other type. The precise definition of “closer” in the proof of such a result will depend on the scoring rule  $S$ . For instance, with the logarithmic scoring rule, assumption A7 can be replaced by the assumption that the relative entropy between the best predictions of two individuals with the same type is smaller than the relative entropy between the best predictions of two individuals with different types. As long as the expectations of respondents are clustered in this sense, truth-telling will remain strictly incentive-compatible. In Subsection 4.7.3, we formalize this point.

Our method further allows that only a subsample of the respondents answer both the MCQ and the prediction question, while the remaining respondents answer the MCQ only. To accommodate this option theoretically, we redefine the matching trigger such that for each answer choice  $i \in A$  there is a respondent  $s \neq r$  who answers  $i$  *and* submits a prediction. Respondents who only answer the MCQ will receive the score  $\bar{S}$ , provided that the (redefined) event  $\mathcal{M}^r$  occurs. This simplifies the mechanism for those respondents without losing incentive compatibility.

Finally, by using other types of scoring rules one can adjust choice-matching to respondents who are not risk-neutral or do not even maximize expected utility (Offerman et al., 2009). Whichever method is chosen, choice-matching adds minimal difficulty to these procedures, since it relies on the same scoring principle for rewarding answers to the MCQ as for rewarding predictions. A respondent who understands the elicitation method will thus easily understand choice-matching incentives as well.

## 4.4 Choice-Matching Generalized

Making predictions about other respondents' answers is an attractive default for our auxiliary task. However, there are situations in which its use cannot be recommended. First, it may be that the stochastic relevance of individual answer types is weak since the distribution of types in a population is well known, and the planner does not want to include another question in the survey to reinstate stochastic relevance. Second, some respondents may have trouble understanding the payments made according to proper scoring rules. Finally, in some situations respondents could expect that predictions of some other respondent type might be more accurate. For example, if a survey asks about occupation, a respondent could have an incentive to claim to be someone that has specialized knowledge of the empirical distribution, for instance a labor economist.

In this section, we show that there is a general principle behind choice-matching which

can be employed by methods that do not rely on predictions.

#### 4.4.1 General Mechanism

To formalize the underlying principle of choice-matching, we first introduce real-valued utility-functions  $u_k(y^r, x^{-r}, y^{-r})$  for  $k \in A$  that depend on all variables except a respondent's reported type  $x^r$ . Since respondents do not necessarily report a prediction, we refer to  $y^r$  in the following as  $r$ 's *y-response*, which may take values in some general response set  $\Omega$ . In the prediction-based model of Section 4.3, the utility functions correspond to the (realized) prediction score:

$$u_k(y^r, x^{-r}, y^{-r}) = \begin{cases} S(y^r, x^{-r}) & \text{in the case of } \mathcal{M}^r \\ 0 & \text{otherwise.} \end{cases}$$

Note that in Section 4.3 we implicitly assumed that the planner knows the respondents are risk-neutral. However, in general, choice-matching does not require such knowledge. What is needed is that the reward system used by the planner induces a game in which different strategies are optimal for different types.

**Definition 21.** Let  $G$  be a (Bayesian) game given by the collection of  $N$  respondents, a set of types  $A$ , a set of potential  $y$ -responses  $\Omega$ , a prior  $P$  and utilities  $\{u_k\}_{k \in A}$ . The game  $G$  is *type-separating* if there is a profile  $\sigma$  such that for every respondent  $r$  and every  $k \in A$ :

- (i)  $\sigma_x^r(t^r) = t^r$
- (ii)  $\sigma_y^r(t^r) = y^{*k}$  for some  $y^{*k} \in \Omega$  if and only if  $t_k^r = 1$  and
- (iii)  $E[u_k(y^{*k}, t^{-r}, \sigma_y^{-r}(t^{-r})) \mid t_k^r = 1, \mathcal{E}^r] > E[u_k(y, t^{-r}, \sigma_y^{-r}(t^{-r})) \mid t_k^r = 1, \mathcal{E}^r]$  for any  $k \in A$ ,  $y \in \Omega$ ,  $y \neq y^{*k}$ .

In words, in a type-separating game there is a profile  $\sigma$  in which respondents declare their types honestly (condition (i)), and in which respondents of identical types give the

same  $y$ -response and respondents of different types give different  $y$ -responses (condition (ii)). Furthermore, this profile is a Nash equilibrium (condition (iii)). Importantly, this equilibrium is not strict in  $x^r$ , since  $u(y^r, x^{-r}, y^{-r})$  does not depend on  $x^r$ , so that the game  $G$  alone is not sufficient to reveal types. For instance, in the setting discussed in Section 4.3, it would not be sufficient to only score respondents on their own predictions, since then they would have no incentive to report their honest answer to the MCQ.

In Section 4.3, conditions (i) – (iii) correspond to separation in the prediction task, which – under assumptions A4-A7 – can be achieved by choosing a strictly proper scoring rule. The idea of the next proposition is that prediction scoring can be replaced by any game that satisfies conditions (i) – (iii).

**Proposition 6.** *Let  $G = \langle N, A, \Omega, \{u_k\}_{k \in A}, P \rangle$  be a type-separating game. Under assumptions A4-A5, any payment rule is strictly incentive compatible if it induces a game  $\langle N, A, \Omega, \{V_k\}_{k \in A}, P \rangle$  in which on event  $\mathcal{M}^r$ :*

$$V_k(x^r, y^r, x^{-r}, y^{-r}) = \lambda u_k(y^r, x^{-r}, y^{-r}) + (1 - \lambda) \bar{u}_k(x^r, x^{-r}, y^{-r})$$

where  $\lambda \in (0, 1)$  and  $\bar{u}$  is the average utility value achieved by the respondents other than  $r$  who submit  $x_s = x_r$ , i.e.,

$$\bar{u}_k(x^r, x^{-r}, y^{-r}) = \frac{\sum_{s \neq r} x^r \cdot x^s u_k(y^s, x^{-s}, y^{-s})}{\sum_{s \neq r} x^r \cdot x^s},$$

and, on the complement of  $\mathcal{M}^r$ :

$$V_k(x^r, y^r, x^{-r}, y^{-r}) = 0.$$

*Proof.* The proof is given in Section 4.7. It is almost identical to the special case in Section 4.3. □

### 4.4.2 Example and Applications

To see how choice-matching can be applied when the second question is not a prediction, consider again the example from Section 4.2 in which each respondent gives a rating for a trial product. As an alternative to scoring respondents on predictions, the company could use the following procedure: first, the company presents a list of existing products to each respondent and each respondent chooses the product on the list she likes most. Again, we use the matching trigger: if there is a product that is not chosen by anyone, then  $r$  receives nothing, or some pre-defined, fixed participation fee. Otherwise,  $r$  participates in a “product lottery”: she receives the product she chose with probability  $\lambda$ , or else she receives the product chosen by a respondent randomly selected among those who give the same rating.

Intuitively, these incentives work if respondents expect that those who have the same assessment of the trial product will have (approximately) the same preference with regard to the products on the list, which is also what the type-separation assumption in proposition 6 entails.<sup>6</sup>

Experimental economics is another promising application domain for generalized choice-matching. Suppose that an experimenter is interested in the willingness-to-pay for a good that cannot be offered for sale, perhaps because its price exceeds the research budget or because it is hypothetical in nature. As the auxiliary task, the experimenter can then ask for the willingness-to-pay for a good which can be incentivized, and link the hypothetical willingness-to-pay to the stated willingness-to-pay demonstrated in the lab. Choice-matching also makes it possible to incentivize survey responses about behavior outside the lab, by linking the respondent’s reward to the one obtained by respondents who give the same answers to the survey questions. For instance, in an experiment on decision-making under risk, respondents often choose among a number of gambles, with one of the gambles

---

<sup>6</sup> The type-separation assumption can be relaxed in a similar fashion as the impersonal updating assumption in Subsection 4.3.1.

randomly chosen to be played out. Using choice-matching, the experimenter could ask a survey question related to risk attitude, e.g., which insurance a respondent has bought. She could then provide incentives for this survey question by paying respondents with some probability according to the gambles chosen by respondents with the same survey answer.

## 4.5 Comparison with Existing Methods

Incentives for non-verifiable MCQs based on predictions of the answer distribution were introduced by Prelec (2004), through the Bayesian Truth Serum (BTS) mechanism. Under BTS, a reported answer receives a high score if its actual frequency exceeds predicted frequency (and predictions are scored with the logarithmic version of  $S$ ). The intuition here is that a respondent of, say, type  $k$  expects that other types will underestimate the frequency of  $k$ -types in the population, making an honest  $k$ -report optimal ex-ante. The exact BTS scoring formula is quite opaque and its theoretical guarantee holds only in the large sample limit (Prelec 2004, Cvitanic et al., 2020). However, the scores deliver a bonus property, in that they also reflect respondent expertise. Precisely, if types are i.i.d conditional on an underlying state of the world, then equilibrium BTS scores rank types according to their posteriors on the actual world state. In principle, one can back out the true world state by focusing on the highest-scoring respondents, and thus improve on majority voting and other crowd-wisdom algorithms (Prelec et al., 2017).

Finite sample incentive-compatibility is obtained by several recent mechanisms, typically under weaker assumptions than needed for BTS. For binary questions, these include the Robust Bayesian Truth Serum (RBTS) (Witkowski and Parkes, 2012b) and Baillon’s (2017) “Bayesian Market.” The latter translates reported types and predictions into the buying and selling of securities whose value is defined by the distribution of reported answers. Some respondents may find the competitive market trading setting more engaging



than scoring rules. For general MCQs, Radanovic and Faltings (2013) is mathematically simpler than both BTS and RBTS, but requires that the highest expected frequency of each type is held by the type itself, i.e,  $p_k^{r,k} > p_k^{r,\ell}$ , for each  $k, \ell \in A$  with  $k \neq \ell$ .<sup>7</sup> The mechanism of Zhang and Chen (2014) involves a two-stage game where respondents revise their predictions after receiving the answer of another respondent. Survey implementation would therefore require contacting respondents more than once.

The (non-parametric) Divergence-based Bayesian Truth Serum (DIV) of Radanovic and Faltings (2014) is closest to (prediction-based) choice-matching. The basic idea is to penalize disagreement in predictions among respondents reporting the same type. A theoretical disadvantage of DIV relative to choice-matching, is that it allows non-honest symmetric strict equilibria that also payoff dominate honest ones. A practical disadvantage is that with DIV each respondent must provide a prediction.

Several mechanisms get by without asking for predictions,. The “peer prediction” method of Miller et al. (2005) assumes that the planner knows the common prior, and is in effect able to compute the posteriors that ideal respondents would supply. More recently, data-intensive methods have attempted to estimate the prior from distributional assumptions or by machine learning (Radanovic and Faltings, 2015, Radanovic et al., 2016, Shnayder et al., 2016, Liu and Chen, 2017, Agarwal et al., 2020). These methods are suitable for settings where large quantities of data are collected on similar questions (e.g., different product ratings).

## 4.6 Conclusion

We have proposed a new way to elicit honest answers to a multiple choice question when honesty cannot be verified, and the planner functions as an outsider, agnostic about the distribution of answers in the population. The method asks respondents to engage in

---

<sup>7</sup>This assumption for instance rules out the common knowledge case of stochastic relevance mentioned in footnote 3.

a scored auxiliary task, which in our canonical version is to predict the distribution of answers in the sample. By answering the MCQ, respondents pool with other respondents endorsing that same answer, and share in their success or failure in the auxiliary task. It is not necessary for everyone to participate in the auxiliary task; the minimal number is just two plus the number of possible answers in the MCQ.

Compared to alternatives in the literature, our mechanism may be easier to explain. Previous empirical tests of the BTS (John et al., 2012 and Weaver and Prelec, 2013) used the so-called “intimidation method” (Frank et al. 2017), where respondents are told that they do not need to understand how the payment formula works, only that it is in their interest to be truthful. Although such black-box instructions do affect the aggregate answer distribution, it is not clear what fraction of respondents find the claim credible.

The specific prediction-based method may not be effective when there is strong public information about the distribution of types, or when respondents of the same type, as defined by the MCQ, have very different information about this distribution. The choice-matching principle can still be applied provided there is some task that induces separation among the honest answers. The planner only needs to find a task for which a separating equilibrium exists, she does not need to know the strategies that constitute this equilibrium.

While surveys have played a major role as a research tool in other social sciences, economists have traditionally been suspicious of stated preferences and beliefs, since there are no consequences for responding carelessly or dishonestly. These unverifiable variables may often be the key variables of social science interest. By linking stated and revealed preferences our method erases, in principle, the methodological boundary between those two types of data.

## 4.7 Additional Material

### 4.7.1 Proof of Proposition 6

Recall the statement of Proposition 6:

**Proposition.** *Let  $G = \langle N, A, \Omega, \{u_k\}_{k \in A}, P \rangle$  be a type-separating game. Under assumptions A4-A5, any payment rule is strictly incentive compatible if it induces a game  $\langle N, A, \Omega, \{V_k\}_{k \in A}, P \rangle$  in which on event  $\mathcal{M}^r$ :*

$$V_k(x^r, y^r, x^{-r}, y^{-r}) = \lambda u_k(y^r, x^{-r}, y^{-r}) + (1 - \lambda) \bar{u}_k(x^r, x^{-r}, y^{-r})$$

where  $\lambda \in (0, 1)$  and  $\bar{u}$  is the average utility value achieved by the respondents other than  $r$  who submit  $x_s = x_r$ , i.e.,

$$\bar{u}_k(x^r, x^{-r}, y^{-r}) = \frac{\sum_{s \neq r} x^r \cdot x^s u_k(y^s, x^{-s}, y^{-s})}{\sum_{s \neq r} x^r \cdot x^s},$$

and, on the complement of  $\mathcal{M}^r$ :

$$V_k(x^r, y^r, x^{-r}, y^{-r}) = 0.$$

*Proof.* As in the proof of Proposition 5, observe that  $r$  cannot influence the choice-matching trigger and should thus condition her expected payoffs on being choice-matched. Since  $G$  is type separating, condition (iii) from definition 4 guarantees that the payment rule is strictly incentive compatible in  $y$ . Consider then the difference in the expected score for respondent  $r$  between reporting  $t^r$  honestly and deviating by making some dishonest report  $x^r$  with  $x_i^r = 1$ :

$$P(\mathcal{E}^r \mid t_k^r = 1) \times (1 - \lambda) E \left[ \bar{u}_k(t^r, x^{-r}, y^{-r}) - \bar{u}_k(x^r, x^{-r}, y^{-r}) \mid T^r = k, \mathcal{E}^r \right]$$

Due to the construction of  $\bar{u}(x^r, x^{-r}, y^{-r})$ :

$$\bar{u}_k(t^r, x^{-r}, y^{-r}) - \bar{u}_k(x^r, x^{-r}, y^{-r}) = u_k(y^{*k}, x^{-r}, y^{-r}) - u_k(y^{*i}, x^{-r}, y^{-r})$$

for  $i \neq k$ . Again, using condition (iii) from definition 4, we have that for  $r$ 's conditional expectation of this payoff difference:

$$E[u_k(y^{*k}, x^{-r}, y^{-r}) - u_k(y^{*i}, x^{-r}, y^{-r}) \mid T^r = k, \mathcal{E}^r] > 0$$

which gives the required result.  $\square$

### 4.7.2 Stochastic Relevance Under Conditional Independence and Identical Distribution

Here we show that our stochastic relevance assumption A6 is satisfied in the common setup in which there is an unknown state of the world conditioning on which respondent types are independently and identically distributed.<sup>8</sup>

We first derive conditions which guarantee stochastic relevance without a matching trigger. These conditions are always satisfied when the state of the world has a continuous distribution, and is generically satisfied when it has a discrete distribution. In the second step, we show that our matching trigger adds only minor additional requirements.

**Stochastic relevance with respect to the state of the world** As before, let  $M > 1$  be an integer and  $A = \{1, \dots, M\}$ . We further write  $\Delta^M$  for the  $M$ -dimensional simplex. We start by considering a single respondent, whose type we denote  $T$ . Suppose that  $\Omega$  is a random vector taking values in  $\Delta^M$ , representing the state of the world, and write

---

<sup>8</sup>Note that assumption A7 is then also satisfied.

$$P(T = k | \Omega = \omega) = \omega_k \quad (4.7.1)$$

for all  $k \in A$ . In the following, we assume that  $\Omega$  is a continuous random variable or a discrete random variable, and we denote by  $P(\omega)$  its probability density/mass function. In this Section, we slightly abuse notation by writing  $T = k$  instead of  $T_k^r = 1$  to mean that respondent  $r$ 's type is  $k$ .

**Definition 22.** *Stochastic relevance w.r.t. to  $\Omega$  holds if for all different  $k, j \in A$ :*

$$E(\Omega | T = k) \neq E(\Omega | T = j) \quad (4.7.2)$$

Our first result states:

**Proposition 7.** *Stochastic relevance w.r.t. to  $\Omega$  holds if and only if for all different  $k, j \in A$  and any constant  $\lambda \in \mathbb{R}$ :*

$$P(\Omega_k = \lambda \Omega_j) < 1 \quad (4.7.3)$$

In words, stochastic relevance w.r.t.  $\Omega$  holds if the ratio of any two state components  $\Omega_k, \Omega_j$  is not known with certainty ax ante. This condition is satisfied if  $\Omega$  follows a continuous distribution on  $\Delta^M$  because the subset of  $\Delta^M$  on which  $\Omega_k = \lambda \Omega_j$  has zero measure. If the distribution of  $\Omega$  is discrete, then (4.7.3) is violated only if for all  $\omega$  s.t.  $P(\omega) > 0$ ,  $\omega_k = \lambda \omega_j$ .

*Proof.* We first rewrite (4.7.2). W.l.o.g. take  $k = 1, j = 2$ . For any  $i \in A$ :

$$\begin{aligned} E(\Omega_i | T = 1) &= \int_{\Delta^M} P(\omega | T = 1) \omega_i d\omega = \int_{\Delta^M} \omega_i \frac{P(T = 1 | \omega)}{P(T = 1)} P(\omega) d\omega \\ &= \int_{\Delta^M} \frac{\omega_i \omega_1}{\int_{\Delta^M} P(\omega') \omega'_1 d\omega'} P(\omega) d\omega = \frac{E(\Omega_i \Omega_1)}{E(\Omega_1)} \end{aligned} \quad (4.7.4)$$

And analogously,  $E(\Omega_i | T = 2) = \frac{E(\Omega_i \Omega_2)}{E(\Omega_2)}$ . Thus, stochastic relevance w.r.t.  $\Omega$  holds if and only if for any  $i \in A$ :

$$\frac{E(\Omega_i \Omega_1)}{E(\Omega_1)} \neq \frac{E(\Omega_i \Omega_2)}{E(\Omega_2)} \quad (4.7.5)$$

We next show that (4.7.3) implies (4.7.5) by contradiction. Suppose that (4.7.3) holds but that stochastic relevance w.r.t.  $\Omega$  is violated, i.e. for any  $i \in A$ :

$$\frac{E[\Omega_i \Omega_1]}{E[\Omega_1]} = \frac{E[\Omega_i \Omega_2]}{E[\Omega_2]} \quad (4.7.6)$$

In particular, setting  $i = 1$  and  $i = 2$ , respectively, (4.7.6) implies:

$$E[(\Omega_1)^2] = E[\Omega_1 \Omega_2] \frac{E[\Omega_1]}{E[\Omega_2]}, \quad E[(\Omega_2)^2] = E[\Omega_1 \Omega_2] \frac{E[\Omega_2]}{E[\Omega_1]} \quad (4.7.7)$$

which yields

$$E[(\Omega_1)^2] E[(\Omega_2)^2] = (E[\Omega_1 \Omega_2])^2 \quad (4.7.8)$$

Note that (4.7.8) is an instance of the Cauchy-Schwarz inequality which holds with equality if and only if there is  $\lambda \in \mathbb{R}$  such that

$$P(\Omega_1 = \lambda \Omega_2) = 1 \quad (4.7.9)$$

Thus, we have a contradiction and (4.7.3) indeed implies (4.7.5).

Finally, it is easy to see that (4.7.5) implies (4.7.3). Note that  $P(\Omega_k = \lambda \Omega_j) = 1$  implies  $\lambda = \frac{E(\Omega_2)}{E(\Omega_1)}$ . This immediately yields equation (4.7.6), in violation to (4.7.5).  $\square$

**Corollary 2.** *Stochastic relevance w.r.t. to  $\Omega$  holds if and only if for all different  $k, j \in A$  there is a set  $S \subset \Delta^M$  of possible values of  $\Omega$  with*

$$P(\Omega \in S | T = k) \neq P(\Omega \in S | T = j) \quad (4.7.10)$$

In words, stochastic relevance w.r.t. to  $\Omega$  holds if and only if every type disagrees with all other types about the probabilities of some states of the world.

*Proof.* Consider the sets

$$S_1 = \left\{ \omega : \frac{\omega_k}{E[\omega_k]} > \frac{\omega_j}{E[\omega_j]} \right\}, \quad S_2 = \left\{ \omega : \frac{\omega_k}{E[\omega_k]} < \frac{\omega_j}{E[\omega_j]} \right\}$$

and recall that  $P(\omega | T = k) = \frac{\omega_k}{E[\omega_k]} P(\omega)$ . Thus, for  $i = 1, 2$  we have

$$P(\Omega \in S_i | T = k) = \int_{S_i} \frac{\omega_k}{E[\omega_k]} P(\omega) d\omega$$

and hence  $P(\Omega \in S_i | T = k) = P(\Omega \in S_i | T = j)$  only if  $P(\Omega \in S_i) = 0$ . From Proposition 7, we know that stochastic relevance holds if and only if for all different  $k, j \in A$  and any constant  $\lambda \in \mathbb{R}$ , (4.7.3) holds. It is straightforward to verify that (4.7.3) is equivalent to either  $P(\Omega \in S_1) > 0$  or  $P(\Omega \in S_2) > 0$  (or both). It follows that stochastic relevance holds if and only if  $P(\Omega \in S_1 | T = k) \neq P(\Omega \in S_1 | T = j)$  or  $P(\Omega \in S_2 | T = k) \neq P(\Omega \in S_2 | T = j)$ .  $\square$

**Stochastic Relevance in a Finite Sample Distribution** Consider a finite sample of  $N$  respondents with types  $T^1, \dots, T^N$  which satisfy conditional independence w.r.t.  $\Omega$ , that is, for all  $s \neq r$  and  $i, k \in A$ :

$$P(T^s = i | \Omega = \omega, T^r = k) = P(T^s = i | \Omega = \omega) \tag{4.7.11}$$

and are identically distributed, so that for all  $s, r \leq N$ :<sup>9</sup>

$$P(T^s = i | \Omega = \omega) = P(T^r = i | \Omega = \omega) = \omega_i \tag{4.7.12}$$

---

<sup>9</sup>Note that it is not restrictive that we identify the state  $\omega$  with the probability  $\omega_i$ . Suppose for instance that in the trial product example from Section 4.2, honest ratings are distributed conditionally i.i.d on the quality of the product, represented by  $\Omega$ . We can then simply define a new random vector  $\tilde{\omega}$ , s.t.  $P(T^r = i | \Omega = \omega) = \tilde{\omega}_i$ .

Next, recall that we defined  $\bar{T}_i^{-r}$  as the frequency of respondents whose type is  $i$  in the sample excluding  $r$ .

**Proposition 8.** *Under conditional i.i.d, stochastic relevance w.r.t.  $\Omega$  implies stochastic relevance w.r.t.  $\bar{T}^{-r}$ .*

*Proof.* W.l.o.g let  $r = 1$ . Then, for all  $i, k \in A$ :

$$E \left[ \bar{T}_i^{-1} | T^1 = k \right] = \frac{1}{N-1} \sum_{s=2}^N Pr \left[ T^s = i | T^1 = k \right] = Pr \left[ T^2 = i | T^1 = k \right] \quad (4.7.13)$$

$$= \int_{\Delta^M} P \left( T^2 = i | T^1 = k, \Omega = \omega \right) P \left( \Omega = \omega | T^1 = k \right) d\omega = E \left[ \Omega_i | T^1 = k \right] \quad (4.7.14)$$

Thus, for given  $j, k \in A$ , we get that  $E \left[ \bar{T}^{-1} | T^1 = k \right] \neq E \left[ \bar{T}^{-1} | T^1 = j \right]$  if  $E \left[ \Omega | T^1 = k \right] \neq E \left[ \Omega | T^1 = j \right]$ .  $\square$

**Sufficient Conditions For Assumption A6** We will next state a result which adapts propositions 7 and 8 to our matching trigger. As before, we write:

$$p^{r,k} := E \left[ \bar{T}^{-r} | \mathcal{E}^r, T^r = k \right] \quad (4.7.15)$$

Recall that our assumption A6 states that if  $j \neq k$ , then:

$$p^{r,k} \neq p^{r,j}. \quad (4.7.16)$$

It will further prove useful to use the following identities, where  $\Omega > 0$  means that  $\Omega_k > 0$  for all  $k \in A$ .



**Lemma 2.** *If the types  $T^1, \dots, T^N$  are conditionally i.i.d w.r.t.  $\Omega$  and  $P(\Omega > 0) > 0$ , then for  $s \neq r$  and any  $i \in A$ :*

$$(i) P(T^s = i | \Omega = \omega, T^r = k, \mathcal{E}^r) = P(T^s = i | \Omega = \omega, \mathcal{E}^r) = \frac{1}{N-1} + \frac{N-1-M}{N-1} \omega_i$$

$$(ii) E(\bar{T}_i^{-r} | T^r = k, \mathcal{E}^r) = \frac{1}{N-1} + \frac{N-1-M}{N-1} E(\Omega_i | T^r = k, \mathcal{E}^r)$$

The first identity allows us to conclude that if  $T^1, \dots, T^N$  are conditionally i.i.d w.r.t.  $\Omega$ , then  $T^1, \dots, T^N$  are also conditionally i.i.d w.r.t. to both  $\Omega$  and the matching trigger. The second identity allows us to relate stochastic relevance w.r.t.  $\Omega$  to stochastic relevance w.r.t.  $\bar{T}^{-r}$ .

*Proof.* In the following, w.l.o.g fix  $r = k = 1$ . We first prove identity (i). For  $s \neq 1$ :

$$P(T^s = i | \Omega = \omega, T^1 = 1, \mathcal{E}^1) = \frac{P(T^s = i, \mathcal{E}^1 | \Omega = \omega, T^1 = 1)}{P(\mathcal{E}^1 | \Omega = \omega, T^1 = 1)} \quad (4.7.17)$$

Let  $\theta^{-1}$  be the set of all possible  $t^{-1} = (t^2, \dots, t^N)$  and let  $\theta_{\mathcal{E}^1}^{-1}$  be the set of all  $t^{-1}$  such that 1's matching trigger is in effect. We then get for the numerator:

$$P(T^s = i, \mathcal{E}^1 | \Omega = \omega, T^1 = 1) = \sum_{t^{-1} \in \theta_{\mathcal{E}^1}^{-1}} P(T^s = i, T^{-1} = t^{-1} | \Omega = \omega, T^1 = 1) \quad (4.7.18)$$

$$= \sum_{\substack{t^{-1} \in \theta_{\mathcal{E}^1}^{-1}, \\ s.t. t^s = i}} P(T^{-1} = t^{-1} | \Omega = \omega, T^1 = 1) = \sum_{\substack{t^{-1} \in \theta_{\mathcal{E}^1}^{-1}, \\ s.t. t^s = i}} P(T^{-1} = t^{-1} | \Omega = \omega) \quad (4.7.19)$$

where the first step uses that each part of the sum is zero when  $t^s \neq k$  and the

second step uses conditional independence. Similarly, we get  $P(\mathcal{E}^1|\Omega = \omega, T^1 = 1) = P(\mathcal{E}^1|\Omega = \omega)$  for the denominator. Thus:

$$P(T^s = i|\Omega = \omega, T^1 = 1, \mathcal{E}^1) = \frac{P(T^s = k, \mathcal{E}^1|\Omega = \omega)}{P(\mathcal{E}^1|\Omega = \omega)} = P(T^s = i|\Omega = \omega, \mathcal{E}^1)$$

which is the first part of identity (i). To get the second part, note that due to conditional i.i.d. w.r.t.  $\Omega$  we can rewrite

$$P(T^{-1} = t|\Omega = \omega) = \prod_{s=2}^N \omega_{ts}$$

so that (4.7.19) yields the same expression for all  $s \neq 1$ . Thus, we are allowed to write

$$P(T^s = i|\Omega = \omega, \mathcal{E}^1) = \frac{1}{N-1} \sum_{s'=2}^N P(T^{s'} = i|\Omega = \omega, \mathcal{E}^1)$$

$$= E(\bar{T}_i^{-1}|\Omega = \omega, \mathcal{E}^1) = \frac{1}{N-1} E((N-1)\bar{T}_i^{-1}|\Omega = \omega, \mathcal{E}^1)$$

Let  $Z = (N-1)\bar{T}^{-1} - 1_M$ , where  $1_M$  denotes an  $M$ -dimensional vector of 1s. Note that conditioning on  $\Omega = \omega$  and  $\mathcal{E}^1$ ,  $Z$  is distributed according to a multinomial distribution, with  $N - M - 1$  draws and parameters  $\omega_1, \dots, \omega_M$ , such that:

$$(N-1) E(\bar{T}_i^{-1}|\Omega = \omega, \mathcal{E}^1) = 1 + E(Z_i|\Omega = \omega, Z \geq 0)$$

$$= 1 + (N - M - 1)\omega_i$$

which also shows the second part of identity (i).

To get to identity (ii), observe that for all  $i \in A$

$$E(\bar{T}_i^{-1}|T^1 = k, \mathcal{E}^1) = \int_{\Delta^M} P(\omega|T^1 = k, \mathcal{E}^1) E(\bar{T}_i^{-1}|T^1 = k, \Omega = \omega, \mathcal{E}^1) d\omega \quad (4.7.20)$$

$$= \int_{\Delta^M} P(\omega|T^1 = k, \mathcal{E}^1) \frac{1}{N-1} \sum_{s=2}^N P(T^s = i|T^1 = k, \Omega = \omega, \mathcal{E}^1) d\omega \quad (4.7.21)$$

$$= \int_{\Delta^M} P(\omega|T^1 = k, \mathcal{E}^1) \left[ \sum_{s=2}^N \frac{1}{(N-1)^2} + \frac{N-M-1}{(N-1)^2} \omega_i \right] d\omega \quad (4.7.22)$$

$$= \frac{1}{N-1} + \frac{N-M-1}{N-1} E(\omega_i|T^1 = k, \mathcal{E}^1) \quad (4.7.23)$$

where in (4.7.22) we used identity (i). This proves identity (ii) as well.  $\square$

We will now come to the key result:

**Proposition 9.** *Let  $N-1 > M$  and suppose that*

- (i) *The types  $T^1, \dots, T^N$  are conditionally i.i.d w.r.t.  $\Omega$ ,*
- (ii)  *$P(\Omega > 0) > 0$*
- (iii) *and for all different  $k, j \in A$  and any constant  $\lambda \in \mathbb{R}$  :*

$$P(\Omega_k = \lambda \Omega_j | \Omega > 0) < 1 \quad (4.7.24)$$

*Then, assumption A6 holds.*

*Proof.* Again fix  $r = 1$ . Invoking Lemma 2, we have

$$E(\bar{T}_i^{-1}|T^1 = k, \mathcal{E}^1) = \frac{1}{N-1} + \frac{N-1-M}{N-1} E(\Omega_i|T^1 = k, \mathcal{E}^1)$$

Thus, under assumptions (i) and (ii), A6 holds if and only if  $E(\Omega|T^1 = k, \mathcal{E}^1) \neq E(\Omega|T^1 = j, \mathcal{E}^1)$  for different  $j, k$ .

Note that  $P(\Omega > 0) > 0$  implies  $P(\Omega > 0 | T^r = i) > 0$  for all  $i \in A$ . Under conditional i.i.d., this implies that for any  $i \in A$ :

$$P(\mathcal{E}^1 | T^1 = i) > P(T^2 = 1, \dots, T^{M+1} = M | T^1 = i) \quad (4.7.25)$$

$$= \int_{\Delta^M} P(\omega | T^1 = i) \prod_{k=1}^M \omega_k > 0 \quad (4.7.26)$$

Thus, we can define  $\tilde{P}(\cdot) := P(\cdot | \mathcal{E}^1)$  and  $\tilde{E}(\cdot) := E(\cdot | \mathcal{E}^1)$ . Lemma 1 ensures that if  $T^1, \dots, T^N$  are independent conditional on  $\Omega$ , they are independent conditional on  $\Omega$  when replacing  $P$  and  $E$  by  $\tilde{P}$  and  $\tilde{E}$ . We can thus apply proposition 7 to  $\tilde{P}(\cdot)$  and  $\tilde{E}(\cdot)$ , which yields that assumption A6 holds if and only if for all  $\lambda \in \mathbb{R}$

$$\tilde{P}(\Omega_k = \lambda \Omega_j) = P(\Omega_k = \lambda \Omega_j | \mathcal{E}^1) < 1 \quad (4.7.27)$$

Finally, note that

$$P(\Omega_k = \lambda \Omega_j | \mathcal{E}^1) = 1 - \int_{\Delta^M} P(\omega | \mathcal{E}^1) I_{\{\omega_k \neq \lambda \omega_j\}} d\omega \quad (4.7.28)$$

$$P(\Omega_k = \lambda \Omega_j | \Omega > 0) = 1 - \int_{\Delta^M} P(\omega | \Omega > 0) I_{\{\omega_k \neq \lambda \omega_j\}} d\omega \quad (4.7.29)$$

and that

$$P(\omega | \mathcal{E}^1) = P(\omega | \mathcal{E}^1, \Omega > 0) = \frac{P(\mathcal{E}^1 | \omega, \Omega > 0) P(\omega | \Omega > 0)}{P(\mathcal{E}^1 | \Omega > 0)} \quad (4.7.30)$$

For  $\omega > 0$ , we have that  $P(\mathcal{E}^1 | \omega, \Omega > 0) > 0$ , so that (4.7.30) is strictly larger than 0 if and only if  $P(\omega | \Omega > 0) > 0$ . Therefore, when (4.7.29) does not equal 1, (4.7.28) does not equal 1 either. Thus, assumption (iii) implies (4.7.27) and under (i) and (ii), assumption

(iii) is sufficient to guarantee A6. □

Note that in addition to the requirements of propositions 7 and 8, proposition 9 adds only the additional requirements that the sample is sufficiently large ( $N - 1 > M$ ) and that the linear independence needed for proposition 7 holds also when restricted to  $\Omega > 0$ . Again, given that  $N - 1 > M$ , it is sufficient that  $\Omega$  has a continuous distribution on  $\Delta^M$ . Thus, the only cases that are excluded by our matching trigger are knife edge cases in which  $\Omega$ 's distribution on  $\Delta^M$  is discrete, and in which, given that all types have a strictly positive (conditional) probability, the ratio of at least two types is known. Since this peculiar situation is unlikely to occur in practice, the additional restriction implied by our trigger is a very minor one.

### 4.7.3 Robustness: No common prior

Our main result holds in a model without a common prior  $P$ . Instead, we assume that each respondent  $r$  has probabilistic beliefs represented by probability measure  $P^r$ , with the corresponding expectation operator given by  $E^r$ . Since a Bayesian game requires a common prior, we need to adjust our definition of incentive compatibility which we do in a similar manner as Witkowski and Parkes (2012a), Radanovic and Faltings (2014) and Baillon (2017). We call a rule  $R$  incentive compatible if:

$$E^r \left[ R_{S,\lambda} \left( t^r, y^{r,k}; T^{-r} \right) - R_{S,\lambda} \left( x, y; T^{-r} \right) \mid t_k^r = 1 \right] > 0$$

for any responses  $(x, y) \neq (t_x^r, y^{r,k})$  and each respondent  $r$ . We continue to assume non-degeneracy and stochastic relevance for each probability measure  $P^r$ . In contrast to Subsection 4.3.1 however, it is now possible that  $y^{r,k} \neq y^{s,k}$  for two respondents  $r, s$  (violating impersonal updating). In order to measure the divergence between these posterior expectations, we use a *divergence function*  $d(x, y) : \Delta^m \times \Delta^m \rightarrow \mathbb{R}_+ \cup \{\infty\}$  mapping two probability vectors into the extended positive real line. We require that  $d(x, y) = 0$  if and

only if  $x = y$ , and define:

**Definition 23.** Posterior expectations  $y^{r,k}$  of respondents satisfy *closeness with respect to divergence*  $d$  if:

$$d(y^{r,k}, y^{s,k}) < d(y^{r,k}, y^{s',k'})$$

for any respondents  $r, s$  and  $s'$ , and for all  $k, k' \in A$  with  $k \neq k'$ .

That is, under closeness the posteriors of two respondents with the same type can differ due to different prior assumptions, but they still agree more (as measured by  $d$ ) than two individuals with different types. Put differently, respondents may bring different information about the distribution of types to the survey, but the information about their own honest response dominates differences with regard to the remaining information. We can exploit such a structure via an extension of a relationship between scoring rules and divergence measures studied by Friedman (1983) and Nau (1985):

**Definition 24.** An SPSR  $f(q; j)$  is *effective* with respect to divergence function  $d$  if for probability vectors  $p^1, p^2, q$ :

$$d(p^1, q) \leq d(p^2, q) \iff \sum_i q_i f(p^1; i) \geq \sum_i q_i f(p^2; i)$$

That is, when  $p^1$  is “closer” according to  $d$  to the true probability  $q$  than  $p^2$  is, then the expectation of scoring rule  $f(p, i)$  using true probability  $q$  is higher for  $p = p^1$  than for  $p = p^2$ .

**Proposition 10.** *Suppose that closeness holds with respect to divergence  $d$  and that SPSR  $f(q, j)$  is effective with respect to  $d$ . Then, choice-matching is strictly incentive compatible.*

*Proof.* To simplify the exposition of this proof, we write  $\tilde{E}^r = E^r[\cdot | \mathcal{E}^r, T^r]$  for player  $r$ 's expectation operator conditional on her type and conditional on  $r$ 's matching trigger under honesty. Compared to proposition 5, incentive compatibility for predictions is unchanged. For answers to the MCQ, consider an individual respondent  $r$  of type  $k$  and suppose that all other respondents respond truthfully. Then, the difference of the expected payoff between reporting  $t^r$  and deviating to  $x^r$  with  $x_{k'}^r = 1$  for some  $k' \neq k$  is:

$$\begin{aligned} \tilde{E}^r \left[ \bar{S}^{-r}(t^r, t^{-r}, y^{-r}) - \bar{S}^{-r}(x^r, t^{-r}, y^{-r}) \right] &\geq \min_{s, s' \neq r} \tilde{E}^r \left[ \sum_{i=1}^M \bar{x}_j^{-r} \left[ f(y_i^{s,k}, i) - f(y_i^{s',k'}, i) \right] \right] \\ &= \min_{s, s' \neq r} \sum_{i=1}^M \tilde{E}^r [\bar{x}_j^{-r}] \left( f(y_i^{s,k}, i) - f(y_i^{s',k'}, i) \right) \\ &= \min_{s, s' \neq r} \left[ \sum_{i=1}^M y_i^{r,k} f(y_i^{s,k}, j) - \sum_{j=1}^M y_j^{r,k} f(y_j^{s',k'}, j) \right] > 0 \end{aligned}$$

where the last inequality follows since by definition of closeness and the effectiveness relation between  $d$  and  $f$ , the bracketed term must be strictly positive for all respondents  $r, s, s'$ .  $\square$

Friedman (1983) and Nau (1985) characterize scoring rules which are effective with respect to a *metric*  $d$ , including the well-known quadratic scoring and spherical scoring rules. In addition to their contribution, it is easy to verify that the logarithmic scoring rule is effective with respect to  $d$  when  $d$  is defined as relative entropy (which is not symmetric and hence not a metric). Thus, for two plausible divergence measures of posteriors, relative entropy and quadratic distance, the two most commonly used scoring rules can be used to allow a strictly separating NE with choice-matching.





# Chapter 5

## Improving Information Aggregation Through Meta-Cognition

### 5.1 Introduction

Just as the industrial revolution saw division of labor in manufacturing of a previously unknown scale, the ongoing digital revolution has brought about an unprecedented level of division of labor in decision-making. For an ever-increasing number of decisions, we have access to the recommendations of many, sometimes millions, of distant strangers. However, the abundance of such data is mirrored by an abundance of ways how to make use of them: the question of how best to combine the judgments of individual members of a group into one decision has been studied in decision theory already by Savage (1954) and has since been continuously topical in different fields, ranging from statistical decision theory (e.g. Stone, 1961, Genest and Zidek, 1986, and more recently Baron et al., 2014, Satopää et al., 2014), psychology (e.g. Hertwig, 2012, Budecsu and Chen, 2014) and game theory (e.g. Feddersen and Pedersdorfer, 1998) to experimental economics (e.g. Anderson and Holt, 1997, Gillen et al. 2017).

In this chapter, I put methods under the microscope which have, despite the wide-

spanning existing literature, received relatively little attention. These are methods which try to solve the problem of aggregation by making use of *metabeliefs* (i.e., beliefs about beliefs) and specifically beliefs about the beliefs of *others*. A recently developed method in this spirit is the “Surprisingly Popular Algorithm” (SPA) proposed by Prelec et al. (2017). The SPA asks respondents to predict the share of all possible answer options and chooses an option if it is more commonly chosen (i.e. more “popular”) than on average predicted. The idea is to penalize answer options that can be expected to be frequent regardless of their correctness.

While this algorithm was successfully tested empirically, it is an open question whether it constitutes the best way to make use of metabeliefs. Prelec et al. (2017) offer a theoretical justification of their algorithm in terms of a model with infinite respondents, but I show that when we consider finitely many respondents, this justification ceases to hold. Instead, I propose a new method that under the assumption of Bayesian responses would be optimal even with finitely many respondents. Specifically, I propose “self-aggregation” (SELF), which lets the crowd choose how to aggregate the judgments of its members. If the crowd members are capable of making full use of their metabeliefs, these aggregation rules will be better than one chosen by an outsider. The idea is thus to apply the “wisdom of crowds” (Surowiecki, 2005) to the aggregation problem itself.

In addition to the SPA, I contrast SELF – theoretically and in an online experiment – with majority voting and with aggregation methods based on judgments of confidence. The latter have served as the traditional inputs for aggregation approaches in decision theory and are therefore a natural benchmark. SELF, SPA and confidence-based methods further share as a common feature that they can be applied “on the spot”. That is, they take as inputs only the information provided by respondents. This stands in contrast to methods which make use of large amounts of data on the track-record or individual characteristics of different experts (e.g. Budescu and Chen, 2014). The advantage of on-the-spot methods is that they can be applied even if such data is not available. Moreover,

methods that give differing weights to different opinions according to a complex learning process lack transparency and as a result may not be accepted by the experts that are consulted or by the decision-maker who wants to employ them.

In my theoretical analysis, I consider a binary-choice model in which an aggregator asks two questions to respondents about their choice among two different options. To analyze the performance of the different methods, I build on the model by Prelec et al. (2017). While simplifying their model on one dimension – by limiting the number of available options to two – I generalize the setting by adding additional uncertainty about the distribution of judgments. I show that in this extended setting, the SPA is not always the optimal rule but that under the Bayesian assumption made by Prelec et al. (2017), the optimal decision-rule is available to respondents and can be communicated efficiently.

A special case of this model allows me to make a more detailed comparison: The SPA is theoretically the best out of the methods considered in Prelec et al.’s (2017) experiment when the two choice-options are about equally likely to be the correct choice. When the prior is not equal however, the SPA can be almost arbitrarily worse than alternatives when the crowd size is small. Moving on from the assumption of Bayesian responses, I study an anchoring and adjustment model in the spirit of Tversky and Kahneman (1974). I show that the theoretical advantage of SELF over the SPA prevails under anchoring. However, if methods differ in terms of the difficulty they pose to respondents and anchoring is more pronounced for the types of judgments required by SELF, the theoretical performance ranking may reverse.

In the online experiment, I test the empirical accuracy of the model by implementing an environment with the statistical properties of my model. This allows me to compare responses directly to the Bayesian prescription, which leads to a more powerful design for comparing methods than has been used in previous work. I find that the SPA shows the highest performance, followed by SELF. Estimating anchoring parameters, I find that most of the deviation from the Bayesian model can be explained by a stronger tendency to

report values in the mid-range of the answer scale for some methods, as predicted by anchoring. These results show that the meta-cognitive abilities of individuals are challenged by complex methods such as SELF and the SPA. However, despite anchoring effects, responses contain sufficient information to outperform methods based on less challenging questions.

### 5.1.1 Related Literature

A statistical argument for using a large sample of respondents rather than asking a single individual was first formulated by Condorcet (1785) in what is now called the Condorcet Jury Theorem (CJT). Condorcet’s work was re-discovered in the social choice literature by Black (1948) and his classical jury result has led to the development of new conditions under which majority voting (and plurality voting for more than two options), still outperforms individual metrics (List and Goodin, 2001, Dryzek and List, 2003). In a broader context, this is part of a “wisdom of crowd” literature which can be traced back to Galton (1907). In this literature, relatively simple statistics (e.g. the mode, median or mean) of a pool of judgments are demonstrated to outperform the best, or average individual member of the group (e.g. Richardson and Stone, 1971, Treynor, 1987, Herzog and Hertwig, 2011).

In settings in which such simple wisdom of crowds techniques do not perform well, a large body of work exists attempting to correct individual judgments before aggregating them, e.g. Herzog and Hertwig (2009), Hertwig (2012), Fan et al. (2019). This literature is heavily influenced by the study of the calibration properties of confidence judgments and probabilistic reasoning capabilities of humans in the tradition of Kahneman and Tversky (1973; 1982).

In contrast to these specific practical applications, Gneiting et al. (2013), Baron et al. (2014) and Raftery et al. (2005), prove the optimality of specific aggregation methods in a given statistical environment. The setting considered in this chapter is considerably

simpler but makes use of a new type of input which does not rely on individual confidences. Instead, individuals are asked directly for judgments which take into account that there are other individuals who provide judgments as well. Such methods are still quite rare (Steyvers and Miller, 2015). Prelec et al. (2017) suggest a method in which respondents are asked for their predictions of the responses of other individuals. The empirical adequacy of their method has been replicated by Lee et al. (2018).<sup>1</sup> A more general version of the SPA based on Bayesian hierarchical modeling is proposed by McCoy and Prelec (2017).

The next section gives an illustration of the main theoretical idea. Section 5.3 is the formal model, starting with the assumption that respondents give answers in a Bayesian fashion and extending from there to an anchoring and adjustment model. Section 5.4 describes the experimental design and Section 5.5 provides the findings of this experiment. I discuss the relation of these findings to the theoretical model in Section 5.6. Section 5.7 concludes and Sections 5.8 and 5.9 provide additional material.

## 5.2 Illustration: A Jury Decision

The focus in this chapter will be on the simplest possible choice problem, a binary decision. Such a problem was also studied by Condorcet (1785) who considered a jury trying to determine whether a delinquent was guilty or not. In his model, all individual judgments are identically distributed and the correctness of their judgments are independent from each other. Condorcet’s observation was that if each jury member has a probability larger than 50% to make the right decision, the majority of the members will reach the correct decision with near certainty as the jury size increases. Clearly, this model lacks some essential features of real-life decision making, or otherwise large majorities would almost never make mistakes.

---

<sup>1</sup>A correction of this article was posted online which showed that the SPA slightly underperformed the modal prediction of media experts and self-declared highly-knowledgeable respondents.

A natural point of departure from Condorcet’s classic model is to relax the idea that the individual judgments of the jury members are independent from each other. For instance, we can imagine a delinquent who is very well-spoken and who is able to afford highly-paid attorneys. Such a delinquent, whether guilty or not, will tend to convince more jury members. Suppose then that when the delinquent is not guilty, each judge has a probability of 80% to come to the conclusion that the delinquent is in fact not guilty. However, even if he is guilty, it may be that each member will vote to acquit with a large, 65% say, probability. If we suppose further that *a priori* (that is, before the jury members have heard the testimony), it is equally probable that the defendant is guilty or not, we see that each jury member has an (unconditional) probability larger than 50% to make the correct decision in line with Condorcet’s requirement. However, the accuracy of the judgments are not independent from each other: If one of the jury members makes an incorrect judgment, it becomes more probable that another judge also reaches the wrong conclusion. In technical terms, the judgments of the jury member are independent only *conditionally on the correct verdict*. This (seemingly) small difference in assumptions has a stark effect on the accuracy of the jury: under the assumption of conditional independence, at least 65% (approximately) of the jury members will vote in favor of the defendant. If the jury votes according to the majority rule, it will thus always acquit the defendant, even if she is guilty. Thus, since we assumed unconditional probability of guilt to be 50%, the jury will reach the wrong verdict about 50% of the time. In a reversal of Condorcet’s initial result, a large crowd can have a lower “expected accuracy” than each of the individual jury members.

As a way out of such situations, one could try to find out which jury members have particularly high accuracy. In the case of the well-spoken defendant, one could for instance try to identify those members which are less easily fooled by appearance. It is however not easy to measure accuracy in the jury case since the actual “ground truth” of guilt or innocence is, in the overwhelming majority of cases, not known (Sunstein, 2006). As

a replacement, we could try to ask the jury members for their own assessment of their expertise. However, as has long been documented, confidence reports are often not well calibrated (Kahneman and Tversky, 1982, Lichtenstein et al., 1982, Keren, 1991, Johnson and Fowler, 2011). Moreover, as pointed out by Prelec et al. (2017), even if individual confidences were fully Bayesian probabilities, they might not be a suitable input for judgment aggregation. In the case of our jury, those members who vote guilty will always be more confident.<sup>2</sup> What is informative about them, rather than their relative confidence, is that there are surprisingly many of them in case the defendant is indeed guilty.

As an alternative to confidence-weighted measures, the Surprisingly Popular Algorithm of Prelec et al. (2017) therefore asks for a *prediction of* the overall response. The jury members, for instance would predict the responses by their fellow jury members. If more jury members voted guilty than on average predicted, “guilty” should be the verdict of the jury (and “not guilty” otherwise). The underlying intuition is simple: jury members should expect a share of 65%-80% to vote to acquit. When the jury is large, and the defendant is not guilty, about 80% of the jury – a larger share than on average expected – will vote to acquit, whereas when the defendant is guilty only 65% will do so, which is less than predicted on average. Thus, with a large enough jury, the SPA will always make the right verdict.

The first question asked by the model in my chapter is how important the assumption of a large crowd is for the argument just outlined. In the next section, I show that it is indeed a crucial assumption. For a smaller crowd, there can be larger sampling variance, so that for instance in our example a share quite different from 65% could vote to acquit if the defendant is in fact guilty and a share different from 80% could vote to convict if the defendant is innocent, even if all the other simplifying assumptions are true. As a result,

---

<sup>2</sup>This is for pure mechanical, statistical reasons: bayesian updating gives a confidence of  $\frac{80\%}{80+65\%} \approx 55.2\%$  for jury members who vote to acquit and a confidence of  $\frac{(100-65)\%}{(100-65)\%+(100-80)\%} \approx 63.5\%$  for jury members who vote to convict.

the optimal Bayesian voting rule can be quite different to the SPA (I do show however, that if the unconditional probability of guilt is 50%, the SPA will be “close” to the optimal rule). This observation is not restricted to a “small crowd”, but rather concerns a general class of situations with which the SPA may not deal well, namely situations in which the innocence or guilt of the defendant does not fully determine the composition of judgments. For instance, there could be idiosyncratic shocks which determine the votes, other than the evidence laid out in court, or the assumption of conditional independence could be violated due to group deliberation.

Pushing the idea to make use of the Bayesian expectations and the domain-knowledge of the jury involved to the limit, I therefore propose a new method, “self-aggregation”, which asks the jury members a rather demanding question: What is the optimal number of guilty votes that we should use as a threshold to convict the defendant? We will see that under mild assumptions, this method can be communicated as efficiently as the inputs required for the SPA. Of course, it could be that this question is over-demanding to the jury members. The truism under the Bayesian case then becomes an empirical question, which I address in the experiment reported in Sections 5.4-5.5. Before this however, I will make the discussion of the jury example rigorous in the next section.

## 5.3 Model

### 5.3.1 Setup and Notation

To formalize the preceding discussion, suppose there are two answer options, denoted 1 and 0 (like “guilty” and “not guilty” in the jury example) and a set  $\mathcal{I} = \{1, \dots, n\}$  of individuals. Each individual  $i \in \mathcal{I}$  makes a *judgment* on which of the two options is the correct one and potentially makes a *meta-judgment* (defined in 5.3.2). The individuals’ judgment is based on a piece of information  $s_i$ , a *signal*, which is the realization of a random variable



$S_i$ . The *true* (or correct) option is a random variable  $T$  equal to either 0 or 1 and the joint distribution of signals and truth is determined by a probability measure  $P$ .

An *aggregator* makes a choice from  $\{0, 1\}$  upon receiving the individual judgments and meta-judgments. As will be made precise below, her aim is to maximize the probability that her choice is correct. The crucial idea is that the aggregator does not know the probability measure  $P$ , whereas it is, at least tacitly, known to the individuals. However, the aggregator may make assumptions about the set of "admissible" priors, denoted  $\mathcal{P}$ . The prior probability  $P(T = 1)$  is denoted  $\pi$ . Moreover I write for the individual signals:

$$P(S^i = 1 \mid T = 1) = p_i \quad (1)$$

$$P(S^i = 1 \mid T = 0) = q_i \quad (2)$$

where  $p_i > q_i$  for all  $i \in \mathcal{I}$ . (1)-(2) are in the form of a signal detection model (Macmillan, 2002) in which individuals try to detect the truth and the parameter  $p_i$  gives the *hit rate* (the rate of positive signals when the underlying state is true) and the parameter  $q_i$  the individuals' *false alarm rate* (the rate of positive signals when the underlying state is false). I denote random variables such as  $T$  and the signals  $S_1, \dots, S_n$  with capital letters.  $S = (S_1, \dots, S_n)$  denotes the random vector of signals and the average signal  $\frac{1}{n}S_i$  is written  $\bar{S}$ . The realizations of these random variables are written  $t, s_i, s$  and  $\bar{s}$ , respectively. In the next subsection, I define properties of the signal distribution which will play a central role throughout the chapter. After that, the relation of the individuals' responses and the signaling structure is defined.

### 5.3.2 Individual responses and the Aggregator's Objective

All four properties discussed in this subsection are commonly made in the judgment aggregation literature. The first property, which is assumed throughout the chapter, relates the individuals' signal and their posterior belief about  $T$ :

**(Signal Property 1)**  $P$  satisfies **Signal Decisiveness** if for each  $i \in \mathcal{I}$  and  $s_i \in \{0, 1\}$ :

$$P(T = s_i | S_i = s_i) > \frac{1}{2}$$

I will also refer to an individual with a signal  $s_i$  such that  $P(T = 1 | S_i = s_i) > \frac{1}{2}$  as “individual  $i$  believes that  $T = 1$ ”. With this convention, Signal Decisiveness means that upon receiving a signal of 0 or 1, each individual believes that  $T = 0$  or  $T = 1$ , respectively.

**(Signal Property 2)**  $P$  satisfies **Exchangeability** if for any permutation  $\phi(1), \dots, \phi(n)$  of  $1, \dots, n$ :

$$P(S_1 = s_1, \dots, S_n = s_n, T = k) = P(S_{\phi(1)} = s_1, \dots, S_{\phi(n)} = s_{\phi(n)}, T = k)$$

for all realizations  $s_1, \dots, s_n$  and  $k \in \{0, 1\}$ .

In words, exchangeability states that permuting the judgments of individuals should not change the probability of the event. That is, it should not matter which expert holds which opinion. This is surely a fairly demanding property and to some extent it limits the possibility of methods which are based on ‘expert selection, since the individuals’ competences are not differentiated with respect to information they are receiving. However, it is still possible to include an imperfect response as a substitute for individual differences in informativeness. I.e., one could include in the model that individuals differ with respect to how accurately they can report aspects of  $P$ .

Under exchangeability, what matters for the aggregator is the share of respondents who believe that  $T = 1$ , i.e. the average signal  $\bar{S}$ . The realization of this variable, the empirical average  $\bar{s}$  of the realizations  $s_1, \dots, s_n$  is then a *sufficient statistic* for the truth parameter  $T$ . This means that conditioning on the random vector  $(S_1, \dots, S_n)$  does not add information with respect to  $T$ . Formally, the conditional distributions  $P(T | \bar{S})$  and

$P(T|S_1, \dots, S_n)$  are identical. The next property refers to this sufficient statistic:

**(Signal Property 3)**  *$P$  satisfies the Monotone Likelihood Ratio Property (MLRP) if for any vector of realizations  $s'$  and  $s$ , for which the average satisfies  $\bar{s}' \geq \bar{s}$  (that is, there is a larger share of 1-signals in  $s'$ ):*

$$\frac{P(S = s', T = 1)}{P(S = s', T = 0)} \geq \frac{P(S = s, T = 1)}{P(S = s, T = 0)}$$

Under MLRP, when a larger number of respondents receives a signal  $S_i = 1$ , then  $T = 1$  becomes relatively more likely compared to  $T = 0$ . It implies that shifting any individuals' signal from 0 to 1 should not make it *less* probable that  $T = 1$  is true. Crucially, MLRP does not exclude the possibility that even for large values of  $\bar{s}$ , the conditional probability  $P(T|\bar{S} = \bar{s})$  can be very low. In particular, if a majority of respondents receives the signal  $S_i = 1$ , it is still possible that  $T = 0$  is more probable than  $T = 1$ .

**Definition 25.** SP1-SP3 together define the *main model*.

*Remark 2.* Under exchangeability,  $p_i = p$  and  $q_i = q$  for all  $i \in \mathcal{I}$ . Then, SP1 and SP3 together imply the parametric restriction that

$$\frac{p}{q} > \frac{1 - \pi}{\pi} \tag{5.3.1}$$

$$\text{and } \frac{1 - q}{1 - p} > \frac{\pi}{1 - \pi} \tag{5.3.2}$$

That is, when  $S_i = 1$ , the ratio of hit- and false-alarm rate  $\frac{p}{q}$  outweighs the prior odds ratio  $\frac{1 - \pi}{\pi}$  (and analogously when  $S_i = 0$ ). The next property gives rise to a sub-case of the model, the “binomial model”:

**(Signal Property 4)**  $P$  satisfies conditional independence of the signals  $S_1, \dots, S_n$  with respect to  $T$  if

$$P(S^i | S^j, T) = P(S^i | T)$$

for all  $i, j \in I$  with  $i \neq j$ .

**Definition 26.** In the *binomial model*, (SP1-SP4) hold.<sup>3</sup>

I coin this subcase the binomial model because under SP1-4, in each state, individual signals follow a binomial distribution. In my notation, the distribution of signals under  $T = 1$ , denoted  $P(S|T = 1)$ , is a binomial  $(n, p)$ -distribution, and  $P(S|T = 0)$ , the distribution of signals under  $T = 0$ , is a binomial  $(n, q)$ -distribution. The binomial model generalizes the case of binary signals considered in Prelec et al. (2017) and it is the setting of the experiment reported in Sections 5.4-5.5.

### 5.3.3 Individual responses and the Aggregator's Objective

Moving from individual signals to responses, I next define what it means for respondents to be truthful. I denote  $i$ 's judgment by  $X_i$ . The vector  $X = X_1, \dots, X_n$ , the average  $\bar{X}$  and the realizations  $x_i, x$  and  $\bar{x}$  are defined analogously to their signal-equivalents.

**(Judgment Property 1)** Individuals are *truthful Bayesians w.r.t. to  $x$*  if for every individual  $i$ :

$$x_i = \begin{cases} 1 & \text{if } P(T = 1 | S_i = s_i) \geq \frac{1}{2} \\ 0 & \text{else.} \end{cases}$$

Note that with signal decisiveness, truthful answers entail that  $x_i = s_i$ . Assuming (JP1), one can identify judgments and signals and can use notation with  $X$  and  $S$  interchangeably.

---

<sup>3</sup>SP3 follows by assuming SP1, SP2 and SP4, so that it is in fact only necessary to assume these 3 conditions.

Besides asking for the individual judgments about  $T$ , the aggregator also poses a *meta-question* to the individuals, to which the individuals provide a response  $y_i$ , called a *meta-judgment*. The meta-question asks for some information about  $P$ . I express the meta-question as a function  $m$  representing the “correct” Bayesian answer to the question. For instance, it could ask “what is the probability that more than 5 respondents will report a judgment of 1?” which would then be captured by  $P(\bar{X} \geq \frac{5}{n} | S_i)$ . Formally, a meta-question is a function  $m$  mapping the conditional probability measures  $P(. | S_i = 1)$ ,  $P(. | S_i = 0)$  into the  $[0, 1]$  interval. This captures two restrictions often implicit in the literature: First, the respondent cannot be asked about information that is conditional on a different signal than the actual realized one, or about unconditional information in which the realization of the signal would still be unknown. For instance, in the jury example a question that would not be allowed would be “Would you think the defendant was guilty, if you would not find the defendant convincing”? Second, I assume that it is impossible for respondents to communicate the entire posterior distribution. In the spirit of Compte and Postlewaite (2018), I model that individuals are limited in their ability to communicate their knowledge about the statistical setting by restricting the answer space for the  $y$ -answer to  $[0, 1]$ . This can still incorporate confidence judgments, SPA-like predictions and even decision-rules. Aside from these assumptions,  $m$  can be any function which maps the conditional distribution  $P(. | s_i)$  into a response in  $[0, 1]$ . We will see below that under assumptions SP1-SP3, this is sufficiently unrestrictive to allow the aggregator to make full use of the individuals’ information if they are giving Bayesian responses, as defined next:

**(Judgment Property 2)** *Individuals are truthful Bayesians w.r.t. to  $y$  if for every individual  $i$ :*

$$y_i = m(P(. | S_i = s_i))$$

If individuals are truthful Bayesians w.r.t.  $x$  and  $y$ , I will simply say that they are

truthful Bayesians. Alternatives to the assumptions of truthful Bayesians are considered in Subsection 5.3.7.

**The Aggregators's aim** The aggregator's task is to define an *aggregation method*. An aggregation method  $A$  is a pair  $(m, C)$  where  $m$  is a meta-question and  $C : \{0, 1\}^N \times [0, 1]^N \rightarrow \{0, 1\}$  is a *choice-rule* which aggregates the vector of individual judgments and meta-judgments into a single judgment (either 0 or 1). The aggregator's aim is to find an aggregation method that is optimal, given the set  $\mathcal{P}$  of possible probability measures. To compare different aggregation methods, I use the following terminology.

**Definition 27.** An aggregation method  $A = (m, C)$  is *better w.r.t. to  $P$*  than  $A' = (m', C')$ , written  $A \succsim_P A'$ , if

$$P(C(X, Y_m) = T) \geq P(C'(X, Y_{m'}) = T) \quad (5.3.3)$$

where  $Y_m$  is the vector of answers to the meta-question  $m$ .

When responses are truthfully Bayesian,  $Y_m$  is the vector of truthful Bayesian meta-judgments:

$$Y_m = (m(P(\cdot | S_1), \dots, m(P(\cdot | S_n)))$$

Following standard convention, I write  $A \succ_P A'$  if (5.3.3) holds strictly and  $A \sim_P A'$  if it holds with equality. Aggregation method  $A = (m, C)$  is thus (strictly) better under  $P$  than  $A' = (m', C')$  if asking meta-question  $m$  and using choice-rule  $C$  will lead to the correct decision with (strictly) higher probability than asking meta-question  $m'$  and using choice-rule  $C'$ . Ideally, the aggregator would like to find a rule which is optimal in the following sense:

**Definition 28.** An aggregation method  $A = (m, C)$  is *optimal w.r.t. to  $P$*  if

$$A \succsim_P A' \quad (5.3.4)$$

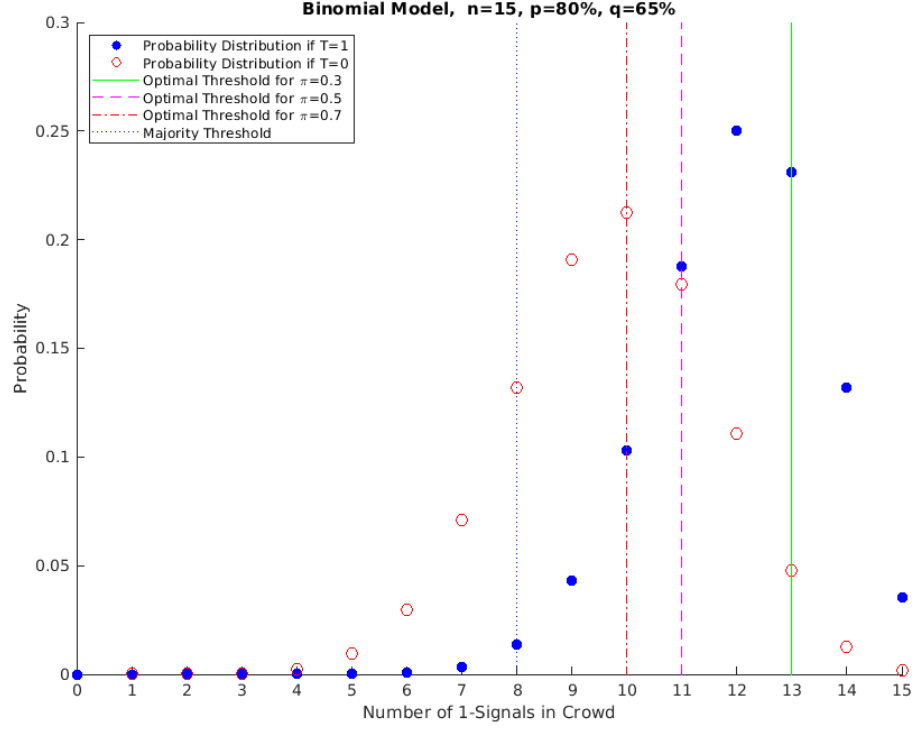


Figure 5.3.1: Optimal threshold values for three different levels of the prior  $\pi$ . The distribution indicated by the unfilled red dots depict the probability of different numbers of individuals to receive a 1-signal if  $T = 0$ , whereas the filled blue dots graph shows the distribution conditional on  $T = 1$ .

for any alternative aggregation method  $A'$ .

As a first observation in the main model, any optimal aggregation method must have a choice-rule of a specific form under Bayesian responses:

**Proposition 11.** *If  $A = (m, C)$  is optimal with respect to each  $P \in \mathcal{P}$ , there exists a function  $t : \mathcal{P} \rightarrow [0, 1]$  such that:*

$$C(x, y) = \begin{cases} 1 & \text{if } \bar{x} \geq t(P) \\ 0 & \text{if } \bar{x} < t(P) \end{cases} \quad (5.3.5)$$

*Proof.* See Section 5.8. □

The basic observation of this proposition is that in the main model, every optimal aggregation method, must partition the  $[0, 1]$  interval into two regions, in one of which the aggregation methods concludes that  $T = 0$  and in one of which it concludes that  $T = 1$ . Figure 5.3.1 gives a graphical intuition of this observation. There, the signaling structure satisfies SP1-SP4 with  $n = 15$ , hit rate  $p = 0.8$  and false-alarm rate  $q = 0.65$ , the parameters from the jury example in Section 2. The picture depicts the optimal threshold-rule for different values of  $\pi$  (the prior probability that  $T = 1$ ). Consistent with MLRP, the ratio of probabilities under  $T = 1$  and  $T = 0$  increases with the number of individuals who receive a signal of 1. Moving from the left to the right, the optimal threshold is at the first point at which the probability ratio exceeds the ratio  $\frac{\pi}{1-\pi}$  of the prior probabilities. At this point, the number of 1-signals is sufficient evidence such that  $P(T = 1|S) > \frac{1}{2}$ . Due to MLRP, all outcomes to the right of this point have an even larger ratio, and are thus even stronger evidence for  $T = 1$ . When  $\pi = \frac{1}{2}$ , the optimal threshold-level (depicted in yellow), is precisely at the point at which the outcome is assigned a higher probability under the distribution conditional on  $T = 1$ . When the prior is lower, the optimal choice rule requires stronger evidence in favor of  $T = 1$  and the optimal threshold (depicted for  $\pi = 0.3$  in green) moves to the right. For a larger prior probability of  $T = 1$ , the optimal required threshold (depicted in unfilled red for  $\pi = 0.7$ ) is lower. We can see however that although the optimal value differs for different values of  $\pi$ , all of the three optimal thresholds are to the right of the majority threshold. Following the majority rule would thus often lead to the incorrect decision.

### 5.3.4 Self-Aggregation, the Surprisingly Popular Algorithm and Confidence-Based Aggregation

**Self-Aggregation** I will next define the different methods in terms of the meta-judgment. The method I propose, SELF-aggregation, asks respondents for which number of 1-judgments,



$T = 1$  becomes more likely than  $T = 0$ . It has the Bayesian answer:<sup>4</sup>

$$m_{SELF}(P(.|s_i)) = \sup_z \left\{ P(T = 1 | \bar{X} = z) \geq \frac{1}{2} \right\}.$$

Under Bayesian responses, the meta-judgments of SELF thus provide the optimal cut-off point  $t(P)$  for the true probability measure  $P$ . The choice-rule of SELF is based on the average  $\bar{y}$  of the responses  $y_i$ :

$$C_{SELF}(x, y) = \begin{cases} 1 & \text{if } \bar{x} \geq \bar{y} \\ 0 & \text{if } \bar{x} < \bar{y} \end{cases} \quad (5.3.6)$$

Since each of the meta-judgments  $y_i$  is equal to the optimal cut-off point  $t(P)$ , SELF is thus an optimal aggregation rule under truthful Bayesian responses. Note however that optimality does not uniquely determine a question and choice-rule. For instance, an aggregation method could ask for an arbitrary transformation of the SELF-question, say the question  $\tilde{m} = m_{SELF} - 1$ , and correct for this transformation in the choice-rule by concluding that  $T = 1$  if  $\bar{x} \geq \bar{y} + 1$ . Or, more sensibly, a different aggregation method could use the geometric mean of the  $y_i$ -judgments, rather than the arithmetic one. However, it seems most natural to state the question as it is done by SELF.

**Surprisingly Popular Algorithm** As was mentioned in Section 5.2, the meta-question in the Surprisingly Popular Algorithm (SPA) proposed by Prelec et al. (2017) asks each individual for their expectation of the average response, i.e.:

$$m_{SPA} = E(\bar{X} | S_i = s_i)$$

where  $E$  is the expectation operator associated with the measure  $P$ . The choice-rule

---

<sup>4</sup>The practical implementation of this question is discussed in Section 5.4.

of the SPA is of the same form as SELF:

$$C_{SPA}(x, y) = \begin{cases} 1 & \text{if } \bar{x} \geq \bar{y} \\ 0 & \text{if } \bar{x} < \bar{y} \end{cases} \quad (5.3.7)$$

**Average Confidence** For the *average confidence* method (AVGCONF), the meta-question asks about the probability that  $i$  assigns to  $T = 1$ , that is:

$$m_{AVG} = P(T = 1 | S_i = s_i)$$

and the choice-rule defined by AVGCONF is:

$$C_{AVGCONF}(x, y) = \begin{cases} 1 & \text{if } \bar{y} \geq \frac{1}{2} \\ 0 & \text{else} \end{cases} \quad (5.3.8)$$

That is, AVG chooses an option if the average respondents believes it to be true with at least 50% probability.

**(Confidence-Weighted) Majority Voting** *Confidence-weighted majority voting* (CONFMAJ) uses the same input as AVGCONF (i.e.  $m_{CONFMAJ} = m_{AVGCONF}$ ), but a different choice-rule:

$$C_{CONFMAJ}(x, y) = \begin{cases} 1 & \text{if } \sum_i x_i y_i \geq \sum_i (1 - x_i)(1 - y_i) \\ 0 & \text{else} \end{cases} \quad (5.3.9)$$

That is, CONFMAJ weighs each individual  $i$  with the probability  $i$  reports that their judgment is correct (since  $y_i$  is the probability  $i$  assigns to  $T = 1$ , the value  $1 - y_i$  is the probability an individual  $i$  with  $x_i = 0$  assigns to this judgment being correct). If the number of individuals who report  $x_i = 1$ , weighted by their reported probability, exceeds the weighted number of individuals who report  $x_i = 0$ , CONFMAJ chooses option 1.

Finally, Majority Voting (MAJ) does not need  $y_i$  altogether and instead uses:

$$C_{MAJ}(x, y) = \begin{cases} 1 & \text{if } \bar{x} \geq \frac{1}{2} \\ 0 & \text{else} \end{cases} \quad (5.3.10)$$

### 5.3.5 Comparisons of Methods with Bayesian Responses

I next compare the performance of different aggregation methods based on the betterness relation defined in Subsection 5.3.3 and under the assumption of truthful Bayesian responses. We can restate a result from Prelec et al. (2017) in the notation that has been introduced thus far.

**Proposition 12.** *[adapted from Theorem 2 in Prelec et al. (2017)] In the binomial model, as  $n \rightarrow \infty$ , if individuals are truthful Bayesians we have*

$$P(C_{SPA}(x, y) = T) \rightarrow 1$$

*Proof.* [sketch] First note that

$$\begin{aligned} \lim_{n \rightarrow \infty} y_i &= \lim_{n \rightarrow \infty} E[\bar{X}|S_i] = E[\bar{X}|T=1]P(T|S_i) + E[\bar{X}|T=0]P(T=0|S_i) \\ &= pP(T|S_i) + qP(T=0|S_i) \end{aligned}$$

So that for each  $i \in \mathcal{I}$ ,  $y_i$  is a strict convex mixture of  $p$  and  $q$ . In the binary case we consider here, the optimality of the SPA then becomes a simple geometric argument.<sup>5</sup> Since  $q < y_i < p$  for each  $y_i$ , we have that  $\lim_{n \rightarrow \infty} P[\bar{X} > \bar{y}|T=1] = 1$  and  $\lim_{n \rightarrow \infty} P[\bar{X} < \bar{y}|T=0] = 1$ , which ensures that the SPA always chooses the correct option.  $\square$

---

<sup>5</sup>This observation is less obvious when there are more than two choice options (the case treated by Theorem 3 in Prelec et al. (2017), which we do not consider here.

The intuition is the same as in the jury example where it was loosely argued that with a large crowd, the share of jury members who vote to acquit will be close to the expected value conditioning on the defendant being truly guilty or innocent.

As argued in Section 5.2, and shown in Figure 5.3.1, the limit result does not hold for majority voting, for instance when  $p = 0.8$  and  $q = 0.65$ . In this case, as  $n \rightarrow \infty$ , we have  $P[\bar{X} > 0.5] = 1$  and the simple majority voting will always choose  $T = 1$  (since respondents whose judgment is 1 will always be in the majority). This will turn out to be the wrong choice whenever  $T = 0$ . A similar point holds for CONFMAJ.<sup>6</sup>

I next show that for a prior  $\pi$  of  $\frac{1}{2}$ , we can make a similar claim to Proposition 12, without relying on limit results.

**Proposition 13.** *In the binomial model given by (SP1-SP4), whenever individuals are truthful Bayesians and  $\pi = \frac{1}{2}$ :*

$$A_{SELF} \succsim_P A_{SPA} \sim_P A_{AVGCON} \succsim_P A_{CONMAJ} \succsim_P A_{MAJ}$$

*for all  $P \in \mathcal{P}$  and for some  $P \in \mathcal{P}$ , each  $\succsim_P$  is strict.*

*Proof.* For the full proof, see Subsection 5.8.2. □

The idea of the proof is as follows: I show first that when respondents are truthful Bayesians, the choice-rules of each of the aggregation methods can be written in the form of (5.3.5). The corresponding values of  $t(P)$  for each of the aggregation methods

---

<sup>6</sup>Note that  $y_i$  has to be at least 0.5 when  $x_i = 1$  and that  $(1 - y_i)$  can be at most equal to 1. Thus, the highest value of  $\bar{x}$  for which CONFMAJ may choose option 0 can be found by setting  $\sum_i 0.5x_i = \sum_i (1 - x_i)$  which can be rewritten as  $\bar{x} = \frac{2}{3}$ . Thus, CONFMAJ will always choose option 1 when  $\bar{x}$  exceeds  $\frac{2}{3}$ , no matter the exact confidence levels reported.

are:<sup>7</sup>

$$t_{MAJ} = 0.5, \quad (5.3.11)$$

$$t_{CONMAJ} = \frac{\frac{1-q}{1-p+1-q}}{\frac{p}{q+p} + \frac{1-q}{1-p+1-q}}, \quad (5.3.12)$$

$$t_{AVGCON} = \frac{p+q}{2}, \quad (5.3.13)$$

$$t_{SPA} = \frac{p+q}{2}, \quad (5.3.14)$$

$$t_{SELF} = \frac{\log\left(\frac{1-q}{1-p}\right)}{\log\left(\frac{1-q}{1-p}\right) + \log\left(\frac{p}{q}\right)} \quad (5.3.15)$$

which are derived from the definitions in (5.3.6)-(5.3.10). We can see that for  $\pi = \frac{1}{2}$ , we have  $t_{AvgConf} = t_{SPA}$ , so the performance of AVGCONF and SPA must be identical. Note further that all the thresholds above – in particular, the optimal one – do not depend on the number of individuals. Intuitively, this is because only the relative evidence in the sample matters, whereas the sample size  $n$  represents the “strength” of the evidence. The latter is only relevant when there is additional, previous information captured by the parameter  $\pi$ .

In the proof, I show that if  $p > 1 - q$ , we have

$$t_{SELF} > t_{SPA} > t_{CONMAJ} > t_{MAJ} \quad (5.3.16)$$

and the reverse holds for  $p < 1 - q$ . An example of such an ordering is depicted in Figure 5.3.2.

Due to MLRP, this ordering of the thresholds relates to an ordering in terms of expected accuracy as follow: The closer the threshold value of a method is to the optimal one (i.e.

---

<sup>7</sup>Here and in the proof in Subsection 5.8.2, I surpress the dependence on  $P$  for simplicity. Since  $P$  is fixed, there should be no ambiguity.

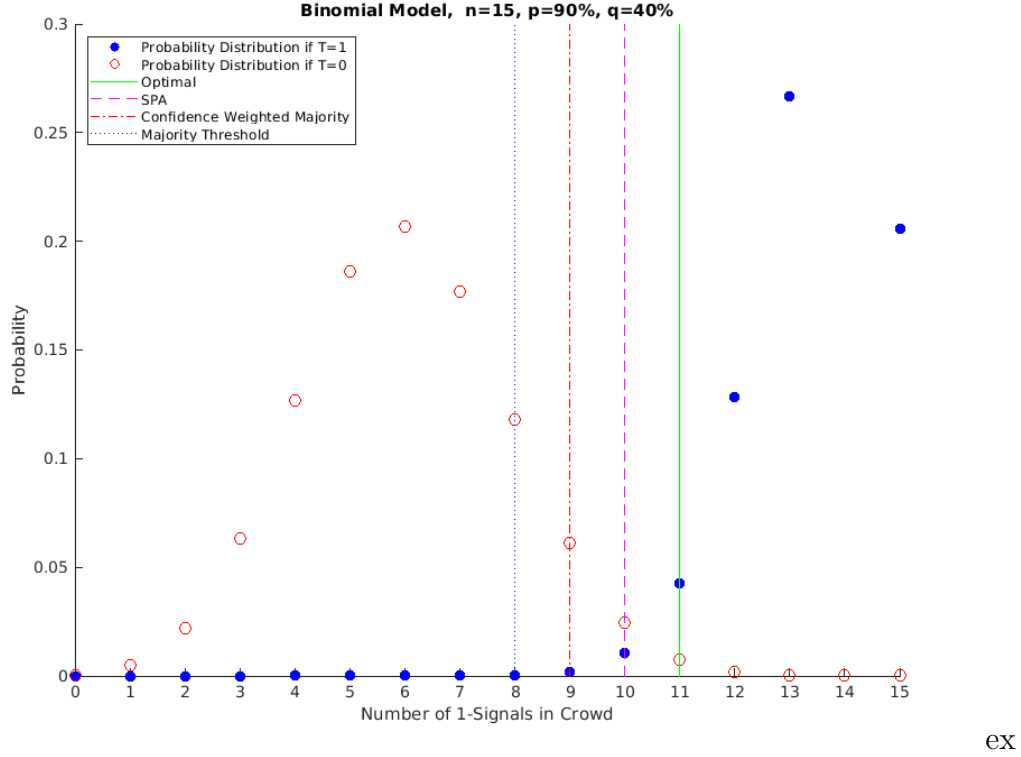


Figure 5.3.2: For  $n = 15$ ,  $p = 0.9$  and  $q = 0.4$  (under an equal prior,  $\pi = 0.5$ ), SELF, SPA, CONFMAJ and MAJ correspond to threshold-rules with thresholds  $t_{SELF} = 11$ ,  $t_{SPA} = t_{AVG} = 10$ ,  $t_{CONFMAJ} = 9$  and  $t_{MAJ} = 8$ , respectively. The ordering of the thresholds from left to right corresponds to an increasing expected performance.

$t_{SELF}$ ), the higher the accuracy. A strict relationship in  $t$  however, does not translate to a strict relationship in  $\succ$ . That is because the empirical average  $\bar{x}$  takes only discrete values  $(\frac{0}{n}, \frac{1}{n}, \dots, \frac{n}{n})$ , so that if for instance  $\frac{8}{n} < t_m < t_{m'} < \frac{9}{n}$  methods  $m$  and  $m'$  have identical expected accuracy.

Therefore, while the relationships among the threshold values is strict, the performance comparison  $\succsim$  in Proposition 13 is not necessarily. However, the example in Figure 5.3.2 shows that all of the relations can be strict at the same time.

### 5.3.6 Shared Information and Asymmetric Prior

In the previous section, the performance of the different methods depended on the fact that  $\pi = \frac{1}{2}$ . I next consider how these considerations change when  $\pi \neq \frac{1}{2}$ . For instance, in the jury example, it could be that there is a commonly assumed base-rate for innocence and guilt in comparable trials. More generally, the prior  $\pi$  can capture the shared information of individuals apart from the private information captured by the signals  $S_i$ . For instance, before a general election respondents in an opinion poll may know the results of previous opinion polls. Or if economists predict the outcome of a replication trial, the shared information could be the outcomes of previous replications. Formally, this shared information can be analyzed via the value of  $\pi$  if the shared information is conditionally independent (w.r.t  $T$ ) of the signal vector  $S$ .

Thus, we can consider the case of differing base rates and shared information by considering how the performance ranking from Proposition 13 changes as the value of  $\pi$  deviates from 0.5. Suppose for instance that  $\pi > 0.5$ . The proof of Proposition 13 in Subsection 5.8.2 also derives the exact values of the thresholds  $t_{SELF}$ ,  $t_{SPA}$ ,  $t_{AVGCON}$ ,  $t_{CONMAJ}$  and  $t_{MAJ}$  for the general case, which are:

$$t_{SELF} = \frac{\log\left(\frac{1-q}{1-p}\right)}{\log\left(\frac{1-q}{1-p}\right) + \log\left(\frac{p}{q}\right)} + \frac{\log\left(\frac{1-\pi}{\pi}\right)}{2N \left[ \log\left(\frac{1-q}{1-p}\right) + \log\left(\frac{p}{q}\right) \right]} \quad (5.3.17)$$

$$t_{SPA} = P(S_i = 1). \quad (5.3.18)$$

$$t_{AVGCONF} = \frac{P(T = 0 | S_i = 0) - 0.5}{P(T = 1 | S_i = 1) + P(T = 0 | S_i = 0) - 1} \quad (5.3.19)$$

$$t_{CONFMAJ} = \frac{P(T=0|S_i=0)}{P(T=1|S_i=1) + P(T=0|S_i=0)} \quad (5.3.20)$$

$$t_{MAJ} = 0.5$$

Rewritten like this, we see that compared to 5.3.15, when  $\pi \neq \frac{1}{2}$ ,  $t_{SELF}$  has an extra additive term of  $\frac{\log(\frac{1-\pi}{\pi})}{2N \left[ \log\left(\frac{1-q}{1-p}\right) + \log\left(\frac{p}{q}\right) \right]}$ . Since  $p > q$ , the denominator of this term is always positive and it is furthermore invariant in  $\pi$ , whereas the numerator is strictly decreasing in  $\pi$ . Thus, the optimal threshold (equal to  $t_{SELF}$  when responses are truthfully Bayesian) is lower the less prior evidence there is since there is less “social” evidence (stemming from the individuals’ signals) needed to conclude that  $T = 1$ . Moreover, since both  $P(T=0|S_i=0)$  and  $P(T=1|S_i=1)$  move closer to the prior, the thresholds for AVGCONF and CONFMAJ similarly decrease.

The threshold  $t_{SPA}$  in turn is increasing in  $\pi$ . For small changes in  $\pi$  around  $\frac{1}{2}$ , this is beneficial because the threshold moves towards the Bayes-optimal one, but it can be harmful to the SPA’s performance for larger changes in  $\pi$ . This dynamic is depicted in Figure 5.3.3, where (for  $n = 100$ ) for most parameters, the SPA is expected to outperform AVGCONF when  $\pi = 0.6$ , but AVGCONF is expected to dominate the SPA when  $\pi = 0.9$ . Since with more extreme priors, less parameters satisfy the parametric restrictions 5.3.1-5.3.2 (or in other words, with extreme priors, the information contained in the prior often dominates the information contained in the individual signals), this dynamic works in favor of the SPA.

Together with Proposition 13, this observation gives a new theoretical explanation for the success of the SPA in a finite sample observed in Prelec et al. (2017) and Lee et al. (2018). On the other hand, the poor accuracy of the SPA in a finite sample for extreme priors, further illustrates that the SPA is not optimal under Bayesian responses. This



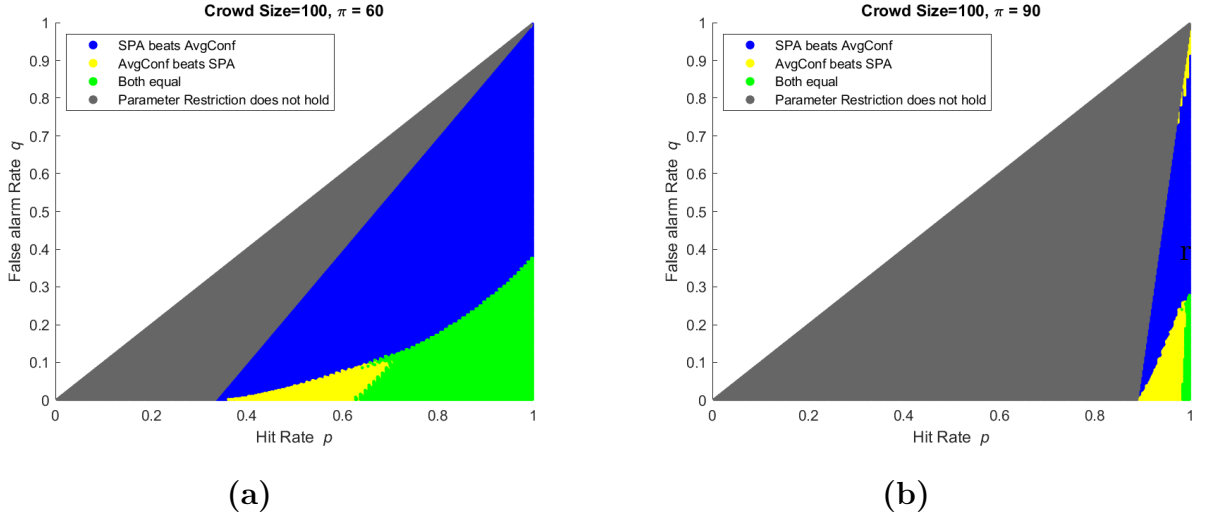


Figure 5.3.3: Theoretically expected performance differences for AVGCONF and SPA for different combinations of the parameters  $p$  and  $q$ .

leaves room for the possibility that other methods may perform better in practice.

### 5.3.7 Comparisons of Methods under Non-Bayesian Judgments

**Anchoring and Adjustment** I will now drop the assumption of Bayesian judgments and consider an anchoring and adjustment model in the spirit of Tversky and Kahneman (1974). Individuals start with an anchor of  $\frac{1}{2}$  and adjust their response towards the Bayesian answer. Individual responses can then be written as:

$$y_i = \frac{\alpha}{2} + (1 - \alpha)m(P|S_i = s_i) \quad (5.3.21)$$

for some  $\alpha \in [0, 1)$ . The larger  $\alpha$ , the more biased are the meta-judgments towards the anchor of  $\frac{1}{2}$  and obviously, when  $\alpha = 0$ , we have no anchoring and meta-judgments are Bayesian again. I exclude the case that  $\alpha = 1$  when meta-judgments are completely uninformative.

The specification in (5.3.21) is deterministic. Therefore, as in Subsection 5.3.4, the meta-judgments of respondents – given their signal – are fixed. Thus, for the comparison

of performances, similar arguments can be used. I write  $t_{\alpha,m}$  for the connected threshold-values of the different methods when meta-judgments are of the form (5.3.21). As a first observation, note that the condition for AVGCONF (given by (5.3.8)) holds with equality if

$$\begin{aligned}\bar{x} &= \frac{0.5 - [\alpha P(T = 1 | S_i = 0) + (1 - \alpha) 0.5]}{[\alpha P(T = 1 | S_i = 1) + (1 - \alpha) 0.5] - \alpha P(T = 1 | S_i = 0) + (1 - \alpha) 0.5} \\ &= \frac{0.5 - P(T = 1 | S_i = 0)}{P(T = 1 | S_i = 1) - P(T = 1 | S_i = 0)} \equiv t_{AVGCONF}\end{aligned}$$

and therefore, for any  $\alpha < 1$ ,  $t_{\alpha,AVGCONF} = t_{AVGCONF}$ . That is, AVGCONF is robust to a distortion in responses of the form (5.3.21), and its expected performance is the same as under Bayesian responses. In contrast, the thresholds of all other methods considered here, move towards the anchor-point of 0.5, and thus reduce their expected accuracy. Taken together, this implies that for larger values of  $\alpha$ , AVGCONF gains in strength relative to other methods. Computations show that for an  $\alpha$ -value of 0.1, SELF and AVGCONF each outperform each other on about half the parameter space. For larger values of  $\alpha$ , AVGCONF starts to dominate SELF for most of the parameter space (see Figure 5.3.4 for an illustration).

Aside from the robustness of AVGCONF, we can prove that also under (5.3.21), SELF is expected to outperform the SPA:

**Proposition 14.** *In the anchoring and adjustment model, if the meta-judgment is as in (5.3.21), then for all  $P \in \mathcal{P}$  and for any  $\alpha \in [0, 1)$ :*

$$A_{SELF} \succsim_P A_{SPA}$$

*and for some  $P \in \mathcal{P}$ , the relation is strict.*

*Proof.* See 5.8.2. □

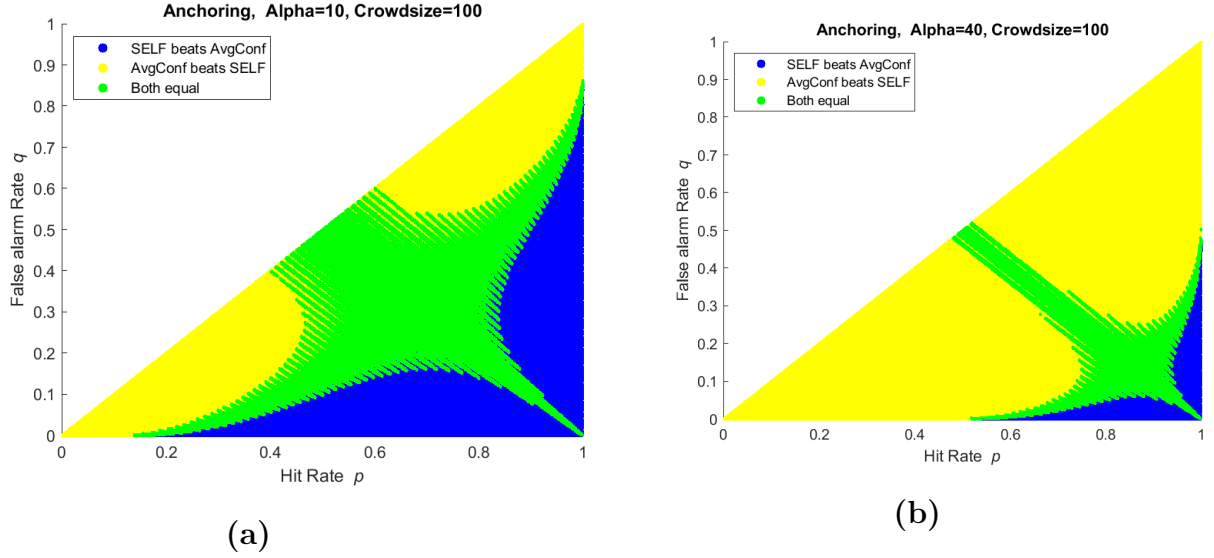


Figure 5.3.4: Theoretically Expected Performance for AVGCONF and SELF under a low (10%, left figure) and middle-sized (40%) anchoring parameter. For low values, the expected performance of AVGCONF and SELF is expected to be fairly balanced on the parameter space, whereas for increasing anchoring, AVGCONF starts to dominate.

Instead of assuming a deterministic and homogenous bias, we could further assume that in the specification (5.3.21), the anchoring parameter depends on the meta-question, i.e. that we have:

$$y_i = \frac{\alpha_m}{2} + (1 - \alpha_m) m(P|S_i = s_i) \quad (5.3.22)$$

In this case, the theoretical performance ordering can change because for some methods, the meta-judgments are closer to the Bayesian prescription. The differences in  $\alpha_m$  could be due to the difficulty connected with giving a particular meta-judgment. For instance, one can suspect that the meta-question posed by SELF is more difficult than the one posed by other methods and that thus the connected anchoring parameter will be higher.

Of course, whether this difference would be sufficiently large to overcome the theoretical advantage of SELF under Bayesian responses is largely an empirical question. The experiment reported in the next sections will shed some light on this question.

## 5.4 Experimental Test

### 5.4.1 Main Task and Treatments

The experimental test of the model in Section 5.3 was performed on the online platform *Prolific*, using the survey software *Qualtrics*. In total, 621 respondents were recruited and divided into six different treatments with approximately 100 respondents each. 598 respondents completed the experiment. To ensure that respondents were able to follow the instructions, the respondent pool was restricted to users with student status and UK or US nationality.

The study follows a  $3 \times 2$ -design depicted in table 5.4.1. The first dimension differs in terms of the meta-question that respondents are asked and the second in terms of the information provided to them. In all treatments, respondents solved 20 inference tasks which are each statistically equivalent to a binary decision in the binomial model with a crowd size of  $n = 5$ . The statistical setting was translated by having two “urns”, Urn A and Urn B, which each contain a number of yellow and blue balls.

In the private information treatments, one of these urns is selected to be the “true” one with equal probability. By clicking on a button, respondents draw a ball from the true urn (see Figure 5.4.1). They are told the color of the drawn ball, but not from which urn it was drawn. The respondents are informed about the number of yellow balls in Urn A (corresponding to the hit rate  $p$  in the binomial model) and in Urn B (corresponding to the false alarm rate  $q$ ). In the shared information treatment, the respondents additionally see a “public draw” that was already made from the urn.<sup>8</sup>

---

<sup>8</sup>For this treatment, it was explained to respondents that the urns – before the public draw is made – are equally likely to be selected. For determining the true urn in the experiment, the conditional probabilities implied by the public draw were used to determine the true urn in each instance.

	Private Information		Shared Information	
	Respondents receive	Respondents provide	Respondents receive	Respondents provide
THRESH	urn properties; a private draw	threshold level	urn properties; private and public draw	threshold level
PRED	urn properties; a private draw	prediction level	urn properties; urn; private and public draw	prediction level
CONF	urn properties; a private draw	confidence level	urn properties; private and public draw	confidence level

Table 5.4.1: The 3x2 design defining the six different treatments of the experiment.

For both private and shared information, the three treatments differ in the type of response the respondents provide, with the different questions corresponding to the different meta-questions discussed in Section 5.3:

1. In the threshold treatments (THRESH), respondents are asked to guess which urn they think their ball was drawn from, depending on the draws of 4 other respondents. Respondents are told that 4 other respondents will draw a ball from the same urn and that they have to make an “adaptive guess”, depending on how many of the other respondents received a yellow or blue ball. The adaptive guess elicits the individual threshold level from Section 5.3.4 in strategic form: For each task, the minimum number of yellow draws<sup>9</sup> needed to make a respondent guess Urn A is interpreted as that respondent’s threshold level. For this number to be well-defined, responses are forced to be monotonic: A respondent can only shift their response from Urn A to Urn B once. Forcing a unique switching point in THRESH has the advantage that a single click was sufficient to fill out the table, which reduces the difference in burden between the three different treatments. Figure 5.9.3 in Subsection 5.9 shows

---

<sup>9</sup>This number includes the respondents’ own signal, since respondents should take this signal into account.

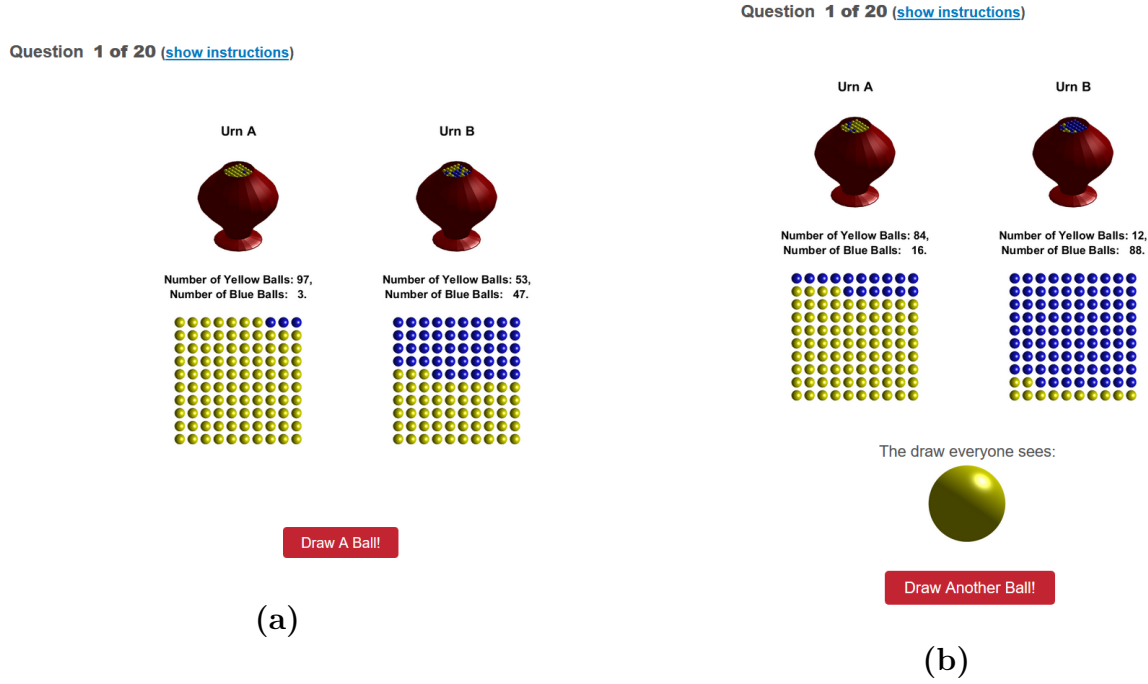


Figure 5.4.1: The interface respondents of all three private information (a) and three shared information treatments (b) see at the beginning of each of the 20 inference tasks. By clicking on the Draw a Ball!-button, respondents receive a signal (a blue or yellow ball) and are prompted to give a response, which differs by treatment.

- an example of an admissible response and provides details on how the interface behaved.
2. In the prediction treatments (PRED), respondents are asked to predict which ball was drawn for four other, randomly determined respondents. To make the type of task comparable to the one performed in THRESH, respondents were asked to select a number of yellow balls ranging from 0 to 4.
  3. In the confidence treatments (CONF), respondents report a probability with which they believe that their ball was drawn from Urn A. This is done via a slider that selects a value from 100 (Urn A for sure) to 0 (Urn B for sure).

For both private and shared information, responses from THRESH are used to evaluate SELF and responses from PRED are used to compute responses for the SPA. Responses

from CONF are used to compute responses for AVGCONF, CONFMAJ and majority voting.

### 5.4.2 Experimental Timeline

All treatments started with the description of the task and an example problem. Furthermore, the payment was explained using an example response. The instructions are followed by the 20 inference tasks during which the full instructions are always available to the respondents. After completion of the tasks, the experiment is concluded with a survey which asks for age, gender and highest attained degree.<sup>10</sup> Respondents were also asked how clear the instructions were and whether they had followed a statistics course (and if yes, when and on which level).

### 5.4.3 Payments

Payments were identical across the private and shared information treatments. In THRESH, after all data were collected, each respondent was matched to 4 other respondents for whom the same urn had been selected for this task. If the guess determined by the draws for these 4 respondents was correct, a respondent received 30 pence for the task (and nothing otherwise). This payment procedure was explained to the respondents before the 20 inference tasks were started. In PRED, the same type of matching as in THRESH was performed after all data were collected. Respondents' predictions of the draws for the 4 other respondents were then scored using a quadratic scoring rule, with  $30 \left[ \bar{x} - y_i^2 \right]$  pence as the payment for each task. Instead of providing this formula, respondents were shown in tabular form how their payment depended on the difference of their prediction and the true number of other respondents who received a yellow ball. In CONF, payments for the

---

<sup>10</sup>The responses to these questions could be complemented and cross-verified with data provided by Prolific.

confidence reports were calculated according to a quadratic scoring rule as well, with

$$\begin{cases} 30 [1 - (1 - y_i)^2] & \text{if Urn A is true} \\ 30 (1 - y_i^2) & \text{if Urn B is true} \end{cases}$$

pence as the payment for each task. As in PRED, and similar to Offerman et al. (2009), respondents were not given the payment formula but were provided with a table that showed the payment for each combination of probability report and true urn. In order to make the table easier to read, confidence reports were rounded to multiples of 5% both in the presentation of the table and in the actual calculation of the payments.

In all treatments, payments were added over all inference tasks.

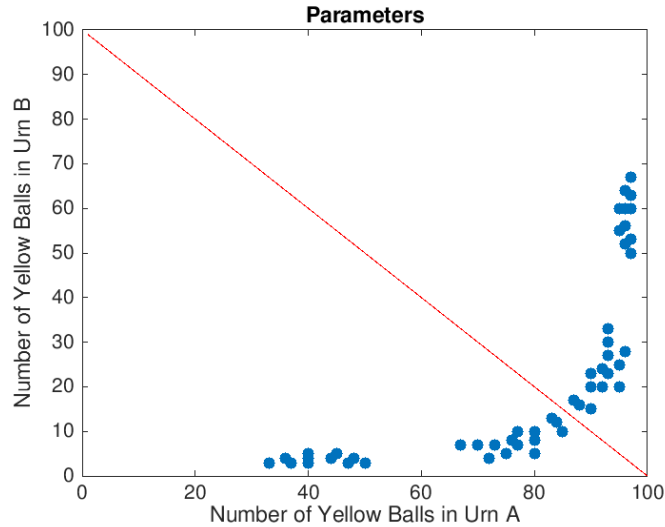


Figure 5.4.2: Scatterplot of the parameters chosen for the private information treatments with the x-axis representing the number of yellow balls in Urn A and the y-axis representing the number of yellow balls in Urn B. In all problems, Urn A contained more yellow balls than Urn B. The plot is symmetric around the diagonal line because of counterbalancing.



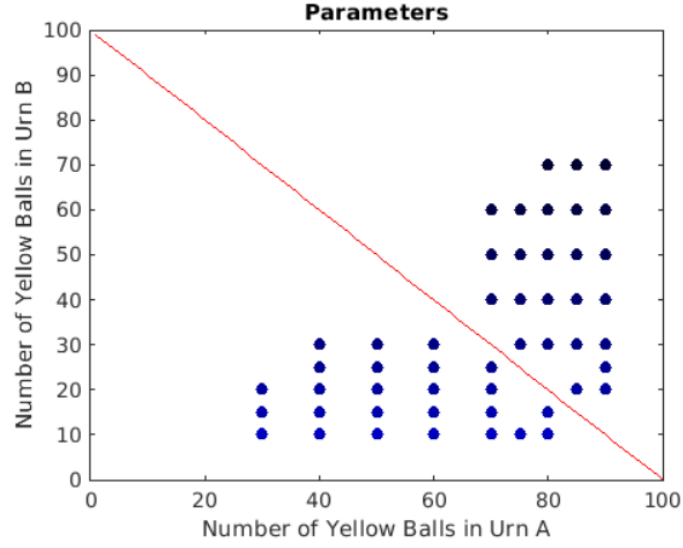


Figure 5.4.3: Scatterplot of the parameters chosen for the shared information treatments.

#### 5.4.4 Parameter Choices

For each respondent, 20 inference tasks were selected from an overall number of 50 problems. In total, each inference task is answered by about 40 respondents per treatment.<sup>11</sup> The order of the 20 tasks is randomized for each respondent individually. The number of yellow and blue balls is counterbalanced.

Parameters were chosen to ensure that they reflect a variety of different theoretical predictions. In particular, in the private information case, the Bayesian prediction for the SPA is the optimal threshold in 20 of the 50 tasks, giving SELF potential to realize its theoretical advantage. Bayesian MAJ and CONFMAJ are optimal for 10 of the 50 tasks, in all of which the SPA is also optimal (which has to be the case, as per Proposition 13). In the shared information treatment, the SPA is optimal for 20 out of 50 tasks, CONFMAJ for 24, average confidence for 1 and majority voting for 30 tasks. The distribution of the parameters is shown in Figure 5.4.2 for private information and in Figure 5.4.3 for shared information.

---

<sup>11</sup>Across treatments, all tasks are answered by at least 37 and at most 42 respondents.

## 5.5 Empirical Results

### 5.5.1 Difference in Performance

The crucial model prediction relates to the performances of Self-Aggregation (SELF), Surprisingly Popular Algorithm (SPA), Average Confidence (AVGCONF), Confidence-weighted Majority Voting (CONFMAJ) and Majority Voting (MAJ). Using the meta-judgments provided by respondents, I estimate the probability  $P(C(X, Y_m) = T)$  on which the performance relation  $\succsim_P$  from (5.3.3) was based in the theoretical section. Denoting by  $\hat{\Pi}_A$  the sum of these probabilities for aggregation method  $A$  over all tasks (from both private and shared information), I find that  $\hat{\Pi}_{SPA} > \hat{\Pi}_{SELF} > \hat{\Pi}_{CONFMAJ} > \hat{\Pi}_{AVGCONF} > \hat{\Pi}_{MAJ}$ . The Bayesian model in contrast predicts the relationship  $m_{SELF} \succsim_P m_{SPA} \succsim_P m_{AVGCONF} \sim_P m_{CONF} \succsim_P m_{MAJ}$  for all probability measures  $P$  given by the parameters chosen for the experiment.

The estimated probabilities are obtained as follows: For every task and every method, I draw random crowds of size  $n = 5$  in the following way: letting  $k$  go from 0 to  $n$ , I randomly select  $k$  respondents who answered the task and received a 1-signal (a yellow ball) and  $5 - k$  respondents who answered the task and received a 0-signal (a blue ball). For each method, the meta-judgments and signals of this crowd determine a decision in favor of Urn A or Urn B. Repeating this procedure for 10,000 randomly composed crowds (thus 50,000 in total for each method-question pair), I obtain an estimate for how often the method will choose Urn A over Urn B when  $k$  respondents receive a 1-signal. Using the parameters of the question, I can calculate from this the probability that the method would choose the correct option.<sup>12</sup> To calculate an “empirical” performance for majority voting, I treat reported probabilities for Urn A larger than 50% as votes for Urn A.

---

<sup>12</sup>Note that the actual true urns that were generated for the respondents are irrelevant for this calculation.

Empirical Performance						
	<b>SELF</b>	<b>SPA</b>	<b>AVGCONF</b>	<b>CONFMAJ</b>	<b>MAJ</b>	<b># Tasks</b>
Private	43.8	44.8	43.6	42.0	41.5	50
Shared	38.5	38.4	35.0	39.4	35.1	50
<b>Pooled</b>	<b>82.3</b>	<b>83.2</b>	<b>78.6</b>	<b>81.4</b>	<b>76.6</b>	<b>100</b>

Bayesian Prediction						
	<b>SELF</b>	<b>SPA</b>	<b>AVGCONF</b>	<b>CONFMAJ</b>	<b>MAJ</b>	<b># Tasks</b>
Private	46.5	44.8	44.8	42.6	41.5	50
Shared	40.6	38.9	37.2	39.4	39.6	50
<b>Pooled</b>	<b>87.1</b>	<b>83.7</b>	<b>82.0</b>	<b>82.0</b>	<b>81.5</b>	<b>100</b>

Table 5.5.1: **Empirical performance in terms of estimated probabilities ( $\hat{\Pi}$ ) compared to the Bayesian prediction.**

Table 5.5.1 displays the estimated probabilities, as well as the predictions of the Bayesian model. It can be seen that while SELF outperforms AVGCONF and majority voting in the pooled data and CONFMAJ in the private information treatments, its performance clearly falls below the one predicted under Bayesian responses. Another large gap between predicted and actual performance can be found in the shared information treatment for majority voting: While under the Bayesian prediction, MAJ is expected to be the second-highest performing method after SELF, it performs substantially worse than all other methods. In the private information treatment however, the empirical performance of MAJ is identical to its Bayesian expectation.

Differentiating the performance between the shared and private information results,

the biggest difference is in the performance of CONFMAJ and AVGCONF. In the shared information treatments, CONFMAJ is the best performing method and performs significantly better than both the SPA and SELF. In the private information treatments, it is however only the 4-th best, performing worse than AVGCONF and only marginally better than MAJ. AVGCONF in turn is the 3rd best method in the private information treatment (on par with SELF), but the worst performing in the shared information treatments.

To test whether the differences in the estimated probabilities (across methods) differ significantly from noise, I conduct a Wilcoxon signed rank test comparing the probabilities for all 100 tasks, yielding a significant (at 5%-level) performance difference for SPA compared to all other methods except CONFMAJ. Furthermore:

$$SPA >^{**} SELF$$

$$\text{and } SELF, CONFMAJ >^{***} AVGCONF >^{***} MAJ$$

where  $>^*$  denotes a difference in performance significant at 10% level,  $>^{**}$  at 5% level, and  $>^{***}$  at 1% level. I show all binary comparisons in table 5.5.2.

### 5.5.2 Response Characteristics

A “sanity check” of the individual responses in CONF is whether respondents, upon receiving a signal favorable for one of the urns (a yellow ball in case of Urn A or a blue ball in case of Urn B), report a probability of at least 50% that the urn is the correct one. In the private information treatment, this is overwhelmingly the case: in 98% of the cases that a respondent received a 1-signal (i.e. drew a yellow ball), a probability of at least 50% that their ball was drawn from Urn A is reported. Correspondingly, 92% of the 0-signals (i.e. a blue ball) led to a probability report of at least 50% that the ball was drawn from Urn B.

	Pooled (597 respondents, 100 tasks)			
	SELF	SPA	AVGCONF	CONFMAJ
SELF	X			
SPA	<b>SPA &gt;(**) SELF</b>	X		
AVGCONF	<b>SELF &gt;(***) AC</b>	<b>SPA &gt;(***) AC</b>	X	
CONFMAJ	CM > SELF	SPA > CM	<b>CM &gt;(***) AC</b>	X
MAJ	<b>SELF &gt;(***) MAJ</b>	<b>SPA &gt;(***) MAJ</b>	<b>AC &gt;(***) MAJ</b>	<b>CM &gt;(***) MAJ</b>
	Private Information (297 respondents, 50 tasks)			
	SELF	SPA	AVGCONF	CONFMAJ
SELF	X			
SPA	SPA > SELF	X		
AVGCONF	SELF > AC	<b>SPA &gt;(***) AC</b>	X	
CONFMAJ	<b>SELF &gt;(**) CM</b>	<b>SPA &gt;(***) CM</b>	<b>AC &gt;(***) CM</b>	X
MAJ	<b>SELF &gt;(**) MAJ</b>	<b>SPA &gt;(***) MAJ</b>	<b>AC &gt;(***) MAJ</b>	<b>CM &gt;(***) MAJ</b>
	Shared Information (300 respondents, 50 tasks)			
	SELF	SPA	AVGCONF	CONFMAJ
SELF	X			
SPA	SELF > SPA	X		
AVGCONF	<b>SELF &gt;(***) CM</b>	<b>SPA &gt;(***) AC</b>	X	
CONFMAJ	<b>CM &gt;(***) SELF</b>	<b>CM &gt;(**) SPA</b>	<b>CM &gt;(***) AC</b>	X
MAJ	<b>SELF &gt;(***) MAJ</b>	<b>SPA &gt;(***) MAJ</b>	AVGCONF > MAJ	<b>CM &gt;(***) MAJ</b>

Table 5.5.2: **Pairwise Comparison with Random Groups of 5.** The table depicts the results of a Wilcoxon Signed Rank Test. The ">" gives the direction fo the test result (e.g., SPA>SELF means that the mean of the signed rank is positive, i.e. favorable to the SPA). Statistical significance is denoted by the (\*)/(\*\*)/(\*\*\*)-superscripts with (\*) denoting significance at the 10% level, (\*\*) at the 5 %-level and (\*\*\*) at the 1%-level. Significant comparisons on any of these levels are highlighted in bold.

As a direct consequence, the simulated decisions for MAJ in the private information case are practically identical to the decisions theoretically predicted by Bayesian responses and the corresponding performance is identical to the expected performance in the Bayesian model. In the shared information treatment however, only 76% of those receiving a 1-signal and 72% of those receiving 0-signal accurately assign a probability of 50% to the urn favored by their signal. This pronounced difference between the two treatments suggests that the additional piece of information (the initial public draw) makes it considerably more difficult for respondents to make accurate probability judgments.

To get a more precise idea of how close responses are to the Bayesian prescription, I estimate the anchoring and adjustment model in Section 5.3.7.

Suggestive evidence for the presence of anchoring can be found in Figures 5.5.1 and 5.5.2 which plot the Bayesian threshold levels against the average threshold reported by respondents in THRESH. We can see that responses are more closely centered around the mid-point of the scale (3) than the Bayesian values. In order to provide an estimate of the anchoring bias, and the corresponding ones for the prediction and threshold levels, one can first rewrite (5.3.22) as:

$$y_i - \frac{1}{2} = (1 - \alpha_m) \left[ m(P|S_i = s_i) - \frac{1}{2} \right]$$

I next run linear regressions on

$$y_{it} - \frac{1}{2} = c_m + (1 - \alpha_m) [m(P|S_{it} = s_{it}) - 0.5] + \varepsilon_{it} \quad (5.5.1)$$

for the three different meta-questions (threshold, prediction and confidence levels)  $m$ . In (5.5.1),  $c_m$  is a constant and  $y_{it}$  is respondent  $i$ 's answer to task  $t$ . The resulting estimates for the  $\alpha_m$  are shown in table 5.5.3, separated for private and shared information. Note that one model prediction of the specification in (5.3.22) is that  $c_m = 0$  for all  $m$ . As can be seen from the regression results, while highly significant, the estimated intercepts

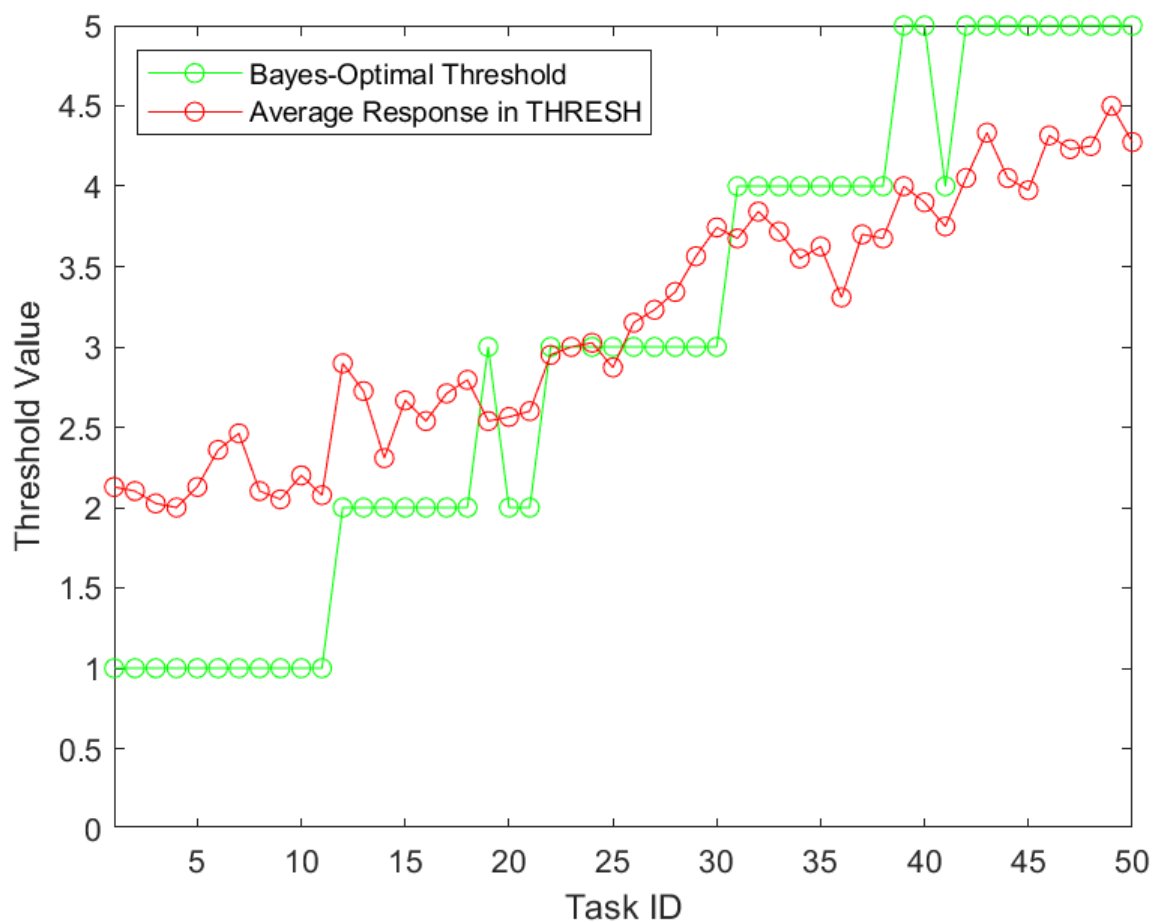


Figure 5.5.1: Optimal threshold levels versus average thresholds provided by respondents, for all 50 tasks in private information treatment. SELF visibly fails to take advantage of tasks such as T1-10 in which the optimal threshold is at an extreme value (1).

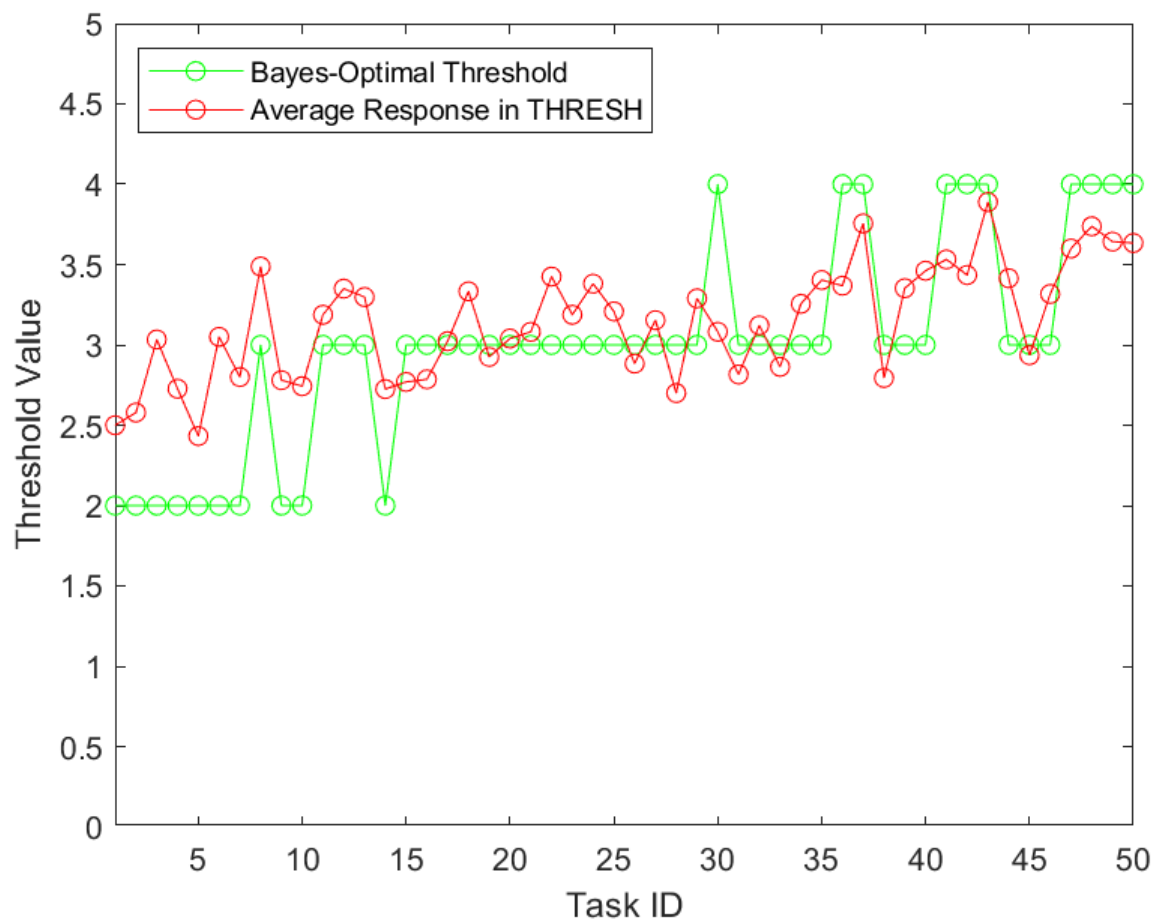


Figure 5.5.2: Optimal threshold levels versus average thresholds provided by respondents in shared information treatment.



in the regressions are small. Only in PRED, there seems to be an additional negative bias, with estimates equalling  $-0.1$  for private information and  $-0.07$  for the shared information treatments. The estimated values further confirm that confidence judgments are considerably more difficult in the shared information environment, as the anchoring parameter estimated for those tasks is much larger (0.33) than in the private information case (0.08).

For both shared information and private information, the ordering of the estimated anchoring parameter is  $\hat{\alpha}_{\text{THRESH}} > \hat{\alpha}_{\text{PRED}} > \hat{\alpha}_{\text{CONF}}$ , suggesting that the confidence question is easiest for respondents to answer. Thus, the anchoring parameters are consistent with a higher performance of the SPA versus SELF, but do not explain why the SPA outperforms AVGCONF in the private information case. The anchoring and adjustment model can thus not fully explain the superior performance of SPA and THRESH compared to AVGCONF.

## 5.6 Discussion

### 5.6.1 Model Adequacy

The model in Section 5.3 predicts most, but not all of the performance relations reported in the previous section. In particular, under Bayesian responses, Proposition 13 should apply in the private information setting. Instead I find  $SPA \succ SELF \sim AVGCONF \succ CONFMAJ \succ MAJ$ . The difference in prediction is compatible with anchoring as discussed in Section 5.3.7, but it cannot give a full account of the data, as AVGCONF performs worse than theoretically expected, even though it is not affected by anchoring in my model. This shows that the theoretical advantage of SELF and SPA over confidence-based methods is robust to behavioral bias, but also that the behavioral processes are more complex than accounted for in Section 5.3.7.

<b>THRESH</b>		
	Private	Shared
$c_{\text{THRESH}}$	0.029(0.00)***	0.029(0.00)***
$\hat{\alpha}_{\text{THRESH}}$	0.56(0.02)***	0.636(0.03)***
$n$ observations	1960	2020
RMSE	0.198	0.218

<b>PRED</b>		
	Private	Shared
$c_{\text{PRED}}$	-0.07(0.01)***	-0.10(0.01)***
$\hat{\alpha}_{\text{PRED}}$	0.34(0.02)***	0.33(0.02)***
$n$ observations	2000	2020
RMSE	0.482	0.211

<b>CONF</b>		
$c_{\text{CONF}}$	0.02	-0.01(0.01)***
$\hat{\alpha}_{\text{CONF}}$	0.08(0.01)***	0.33(0.02)***
$n$ observations	2000	2000
RMSE	0.145	0.227

Table 5.5.3: **Parametric Estimates for Anchoring Parameter.** Ordinary least squares regression on equation (5.5.1). Standard errors in brackets. (\*): significance at the 10% level, (\*\*) at the 5 %-level and (\*\*\*) at the 1%-level.

### 5.6.2 Experimental Design

With respect to the experimental design, the type of task chosen for my experiment merits some additional discussion. The experimental task is of a different type than the ones used in previous work by Prelec et al. (2017) and Lee et al. (2018). Previously, the SPA was tested in tasks where a ground truth was either known to the experimenters or became known to them during the course of the experiment (including trivia tasks, the monetary value of different artwork, and NFL results). For such tasks, it is easy to see for respondents that there is a true and a false answer and therefore experimental results may be expected to have higher external validity with regard to the problems that judgment aggregation methods are ultimately to be applied to.

However, for these tasks it is not possible to know the full distribution of individual judgments. While the conditional distribution under the actual ground truth can be estimated from a large respondent sample, the counterfactual distribution conditioning on the counterfactual cannot be estimated. As a result, previous studies could not compare different methods in their ability to provide the optimal Bayesian answer. This limits comparisons of two aggregation methods to observing whether for a given task one method correctly identifies the ground truth and the other does not. The statistical power of such comparisons is low. For instance, in Prelec et al. (2017), a significant difference between performance of SPA and MAJ was established when pooling the answers across 6 studies, but did not hold for individual studies.

In my experiment, the full distribution of individual judgments is known both under  $T = 1$  and under  $T = 0$  and thus responses translate directly to a probability of choosing the correct option. These probabilities can then be compared per task, which is statistically more powerful. As an illustration of the difference, suppose that responses were fully Bayesian and that we compare the performance of SPA and SELF. Since SELF has a higher probability than SPA of choosing the correct option in 32 tasks (and never has

a lower one) in the private information treatment, we could even fix a significance level as low as  $\alpha = 0.001$  and still power would be 100% (since responses are Bayesian and therefore noiseless). In contrast, when a ground truth is randomly chosen for each task and methods are only scored on detecting this ground truth per task, the performance becomes random, even under noiseless, Bayesian responses. For instance, the statistical power of the Wilcoxon signed rank test for the comparison of SELF and SPA (given the standard significance level of  $\alpha = 0.05$ ) is only around 0.17.

It is possible however, that the statistical framing hurts the performance of SELF somewhat, since it makes the description of the task in THRESH quite complicated. For instance, it requires to explain to respondents that for the four additional draws, the true urn is kept the same. When choosing a task of the type found in Prelec et al. (2017) or Lee et al. (2018), this does not need to be explained since the ground truth is obviously the same for all respondents. This concern is to some degree mitigated by the answers to the final survey at the end of the experiment: The ratings respondents give to the instruction clarity in THRESH are on average only minimally worse than the ones given for CONF and PRED.

### 5.6.3 Possible Model Extensions

The estimates for the anchoring parameters in Section 5.3.7 suggest that confidence-levels, predictions and threshold-levels lie on a scale of methods which are increasingly demanding in terms of respondents' meta-cognition. An interesting avenue for future research is to establish whether there is some optimal level of demandingness that makes maximal use of the information that respondents are able to process.

Furthermore, I have not tapped into the strategic aspects of using SELF in practice when respondents may have an interest in misreporting their judgment. In contrast, the SPA is founded in the incentive literature following Prelec (2004), which has focused on

providing incentives for unverifiable judgment reports and, just like the SPA, made use of predictions by respondents about the judgments of others. It may be an interesting new impulse for the incentive literature to make other meta-cognitive aggregation methods incentive-compatible.

## 5.7 Conclusion

In this chapter, I have first proposed a theoretical model that extends Prelec et al. (2017) by introducing some further uncertainty about the distribution of judgments. In this model, their “Surprisingly Popular Algorithm” (SPA) still outperforms majority voting and confidence-weighted majority voting in expectation, but may not be equal to the optimal aggregation rule an aggregator who knows the whole statistical setting would choose. If respondents have sufficient knowledge about the judgment distribution and are Bayesian reasoners, it is an improvement over the SPA to ask respondents directly for the optimal aggregation scheme, a method I call self-aggregation (SELF). I tested this idea in an online experiment in which I implemented the model directly in a stylized urn setting. The findings suggest that it is difficult for respondents to make full use of the information provided to them, but that there is still a performance gap between the SPA and SELF on the one hand and confidence-based aggregation rules on the other. This shows that while responses are clearly non-Bayesian, there is still a considerable amount of information contained in the responses given by the experiment participants.

With regard to potential applications, it is of course the ultimate aim that new aggregation methods are applied to settings more complex than the binary decision problems I studied. The theoretical properties of these methods may be cumbersome to study and the corresponding questions might be even more difficult for respondents to answer. However, as can be seen by the high performance of meta-cognitive methods in the experiment, the reward may well be worth this extra cost.

## 5.8 Proofs for Section 5.3

### 5.8.1 Proof of Proposition 11

First, fix some  $P$ . Clearly, 5.3.4 is maximized<sup>13</sup> if

$$C_P(x, y) = \begin{cases} 1 & \text{if } P(T = 1 | X = x) > P(T = 0 | X = x) \\ 0 & \text{if } P(T = 1 | X = x) < P(T = 0 | X = x) \end{cases}$$

Next, write  $g(z) \equiv \frac{P(T=1|\bar{X}=z)}{P(T=0|\bar{X}=z)}$ , so that

$$P(T = 1 | \bar{X} = z) \begin{cases} \geq P(T = 0 | \bar{X} = z) & \text{if } g(z) \geq 1 \\ < P(T = 0 | \bar{X} = z) & \text{if } g(z) < 1 \end{cases}$$

MLRP states that  $g$  is increasing in  $z$  and MLRP and signal decisiveness together imply that  $g(1) > 1$  and  $g(0) < 1$ . There must then be a  $z_P$  such that  $g(z) \geq 1$  for all  $z \geq z_P$  and  $g(z) < 1$  for all  $z < z_P$ .

□

### 5.8.2 Proof of Proposition 13

As mentioned in the main text, the main idea of the proof is to write each of the aggregation methods in the form (5.3.5), where I denote the  $t$ -functions corresponding to the different aggregation methods as  $t_{SELF}$ ,  $t_{SPA}$ ,  $t_{AVG}$ ,  $t_{CONFMAJ}$  and  $t_{MAJ}$ , respectively (omitting the dependence on  $P$  for simplicity). I then show that for all  $P$ , either  $t_{SELF} > t_{SPA} = t_{AVG} > t_{CONFMAJ} > t_{MAJ} = \frac{1}{2}$  or  $t_{SELF} < t_{SPA} = t_{AVG} < t_{CONFMAJ} < t_{MAJ} = \frac{1}{2}$ .

As a direct consequence of Proposition 11, choosing  $m(P|s_i) = z_P$  and averaging these

---

<sup>13</sup>The decision can be arbitrary if  $P(T = 1 | X = x) = P(T = 0 | X = x)$ .

responses, as is done by SELF, is optimal. The exact value of  $t_{SELF}$  given in the main text is not needed, but it is derived in 5.8.4 for completeness.

I next show that if  $p > 1 - q$  and  $\pi = \frac{1}{2}$ , then

$$t_{SELF} > t_{SPA} = t_{AVGCONF} > t_{CONFMAJ} > t_{MAJ} = \frac{1}{2} \quad (5.8.1)$$

which is the relation from Figure 5.3.2. The case of  $p < 1 - q$  is analogous. Due to MLRP, a threshold is then better the closer it is to  $t_{SELF}$ . I next prove these relations, starting with (5.8.1) and moving from the right to the left.

**$t_{CONFMAJ} > t_{MAJ}$**  : We have  $t_{MAJ} = 0.5$  by definiton. To derive  $t_{CONFMAJ}$ , note that  $C_{CONFMAJ}(x, y)$  equals 1 if  $\sum_i x_i y_i \geq \sum_i (1 - x_i)(1 - y_i)$ . Replacing  $y_i$  by  $P(T = 1 | x_i)$ , we get that this holds with equality if

$$\bar{x} = \frac{P(T = 0 | S_i = 0)}{P(T = 0 | S_i = 0) + P(T = 1 | S_i = 1)} \equiv t_{CONF}$$

and since  $P(T = 0 | S_i = 0) = \frac{1-q}{1-p+1-q}$  and  $P(S = 1, T = 1) = \frac{p}{p+q}$ ,  $p > 1 - p$  implies that  $P(S = 0, T = 0) > P(S = 1, T = 1)$  so that  $t_{CONF} > 0.5$ .

**$t_{SPA} > t_{CONFMAJ}$**  : For the SPA, under Bayesian responses each respondent reports  $y_i = E[\bar{x} | x_i]$  and therefore  $\bar{x} \geq \bar{y}$  holds if

$$\bar{x} \geq \bar{x} E[\bar{x} | x_i = 1] + (1 - \bar{x}) E[\bar{x} | x_i = 0]$$

Note that the prior expectation  $E[\bar{x}]$  equals  $P(S_i = 1)$ , the unconditional probability that an arbitrary signal  $S_i$  equals 1. We can then write

$$E[\bar{x}] = P(S_i = 1) E[\bar{x} | x_i = 1] + [1 - P(S_i = 1)] E[\bar{x} | x_i = 0],$$

so that the inequality above holds with equality when

$$\bar{x} = P(S_i = 1) \equiv t_{SPA}$$

Thus, when individuals are truthful Bayesians, the implicit threshold associated with the SPA equals the unconditional probability that  $S_i = 1$ . Furthermore, we can rewrite the threshold associated with confidence-weighted majority voting as follows:

$$\begin{aligned} t_{CONF} &= \frac{P(T = 0 | S_i = 0)}{P(T = 0 | S_i = 0) + P(T = 1 | S_i = 1)} \\ &= \frac{\frac{P(S_i=0, T=0)}{P(S_i=0)}}{\frac{P(S_i=0, T=0)}{P(S_i=0)} + \frac{P(S_i=1, T=1)}{P(S_i=1)}} = \frac{\frac{P(S_i=0, T=0)}{P(S_i=0)}}{\frac{P(S_i=0, T=0)P(S_i=1)}{P(S_i=0)P(S_i=1)} + \frac{P(S_i=1, T=1)P(S_i=0)}{P(S_i=0)P(S_i=1)}} \\ &= P(S_i = 1) \frac{P(S_i = 0, T = 0)}{P(S_i = 0, T = 0)P(S_i = 1) + P(S_i = 1, T = 1)P(S_i = 0)}. \end{aligned}$$

If

$$P(S = 0, T = 0) \leq P(S = 1, T = 1),$$

then we can replace  $P(S_i = 1, T = 1)$  by  $P(S_i = 0, T = 0)$  in the numerator, such that

$$\begin{aligned} t_{CONF} &\leq P(S_i = 1) \frac{P(S_i = 0, T = 0)}{P(S_i = 0, T = 0)P(S_i = 1) + P(S_i = 0, T = 0)P(S_i = 0)} \\ &= P(S_i = 1) = t_{SPA}, \end{aligned}$$

with the last equality because  $P(S = 0) + P(S = 1) = 1$ . This establishes that  $t_{SPA} > t_{CONF}$ .

**$t_{SPA} = t_{AVG}$ :** For the average confidence method, we get that the inequality  $\bar{y} \geq \frac{1}{2}$  holds with equality if



$$\bar{x}P(T=1|S_i=1) + (1-\bar{x})P(T=1|S_i=0) = \frac{1}{2} \quad (5.8.2)$$

but also

$$\frac{1}{2} = \pi = P(T=1) = P(S_i=1)P(T=1|S_i=1) + P(S_i=0)P(T=1|S_i=0)$$

$$= P(S_i=1)P(T=1|S_i=1) + [1 - P(S_i=0)]P(T=1|S_i=0)$$

so that equation 5.8.2 solves for  $\bar{x} = P(S_i=1)$  and thus

$$t_{SPA} = t_{AVG} = P(S_i=1)$$

$t_{SELF} > t_{SPA}$ : Finally, note that in the binomial model:

$$P(\bar{S} = \bar{s} | T = t) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } t = 1 \\ \binom{n}{k} q^k (1-q)^{n-k} & \text{if } t = 0 \end{cases}$$

where  $k = n\bar{s}$  is the number of respondents with a 1-signal. We can write the log-likelihood for  $T=1$  and  $T=0$  as

$$l(\bar{s}) = \log \frac{P(\bar{S} = \bar{s} | T = 1)}{P(\bar{S} = \bar{s} | T = 0)} = k[\log p - \log q] + (n-k)\log[(1-p) - (1-q)]$$

An optimal threshold is set such that this log likelihood ratio is positive for all  $\bar{s}$  above the threshold and negative for all lower values. Due to MLRP, it is then sufficient to show that for  $\bar{s} = t_{SPA}$  the log-likelihood ratio is negative. Writing  $p = t_{SPA} + \varepsilon$  and  $q = t_{SPA} - \varepsilon$

where  $\varepsilon = \frac{p-q}{2}$ , we get

$$l(t_{SPA}) = nt_{SPA} \log \left( \frac{t_{SPA} + \varepsilon}{t_{SPA} - \varepsilon} \right) - n(1 - t_{SPA}) \log \left( \frac{1 - t_{SPA} + \varepsilon}{1 - t_{SPA} - \varepsilon} \right) \quad (5.8.3)$$

Since  $x \log \left( \frac{x+c}{x-c} \right)$  is decreasing in  $x$  on  $x > c > 0$ ,<sup>14</sup> we get that (5.8.3) is negative if and only if  $t_{SPA} > 1 - t_{SPA}$ , which is equivalent to  $p > 1 - q$ . This shows that  $t_{SELF} > t_{SPA}$ .

Together, these inequalities establish that  $t_{MAJ} < t_{CONF} < t_{SPA} = t_{AvgConf} < t_{SELF}$ , which gives the weak ordering in terms of performance. The example in Figure 5.3.2 shows strict inequality, concluding the proof.

□

### 5.8.3 Proof of Proposition 14

The idea is to show the following. First, if  $t_{SELF} > t_{SPA} > \frac{1}{2}$ , then clearly

$$t_{SELF} > \alpha t_{SELF} + (1 - \alpha) \frac{1}{2} > \alpha t_{SPA} + (1 - \alpha) \frac{1}{2} > \frac{1}{2}.$$

Thus, by the same idea as in the proof of Proposition 13, we can conclude that  $A_{SELF} \succ A_{SPA}$  if whenever  $p > 1 - q$ :

$$t_{SPA, \alpha} < \alpha t_{SPA} + (1 - \alpha) \frac{1}{2}$$

To see this, I will show that whenever a method  $A$  can be written as:

$$t_A = \frac{m_A(P(.|S_i = 0))}{m_A(P(.|S_i = 1)) + m_A(P(.|S_i = 0))} \quad (5.8.4)$$

---

<sup>14</sup>  $\frac{\partial}{\partial x} x \log \left( \frac{x+c}{x-c} \right) = \log \left( \frac{x+c}{x-c} \right) - \frac{2xc}{(x-c)(x+c)}$ . Differentiating again gives  $\frac{4c^3}{(x-c)^2(x+c)^2} > 0$ . Together with  $\lim_{x \rightarrow \infty} \frac{\partial}{\partial x} x \log \left( \frac{x+c}{x-c} \right) = 0$ , we get that  $\frac{\partial}{\partial x} x \log \left( \frac{x+c}{x-c} \right) < 0$  for all  $x$ .

then, in the anchoring and adjustment model  $t_{A,\alpha} < \alpha t_A + (1-\alpha)\frac{1}{2}$  whenever  $t_A > \frac{1}{2}$  (and the reverse if  $t_A < \frac{1}{2}$ ). This latter fact follows by straightforward calculation. For any numbers  $u, v \in (0, 1)$ :

$$\begin{aligned} & \left[ \alpha \frac{u}{u+v} + (1-\alpha) \frac{1}{2} \right] - \frac{\alpha u + (1-\alpha)\frac{1}{2}}{\alpha(u+v) + 1-\alpha} \\ &= \frac{\alpha}{2} \left[ \frac{u-v}{(u+v)} - \frac{u-v}{[\alpha(u+v) + 1-\alpha]} \right] \propto (u-v)(1-\alpha)(1-(u+v)) \end{aligned}$$

which is positive if and only if  $u+v < 1$ , i.e. if the denominator in the initial expression is smaller 1. Noting next that

$$t_{SPA} = \frac{E[\bar{X} | S_i = 0]}{1 - E[\bar{X} | S_i = 1] + E[\bar{X} | S_i = 0]},$$

we can further observe that  $1 - E[\bar{X} | S_i = 1] + E[\bar{X} | S_i = 0] > 1$  since  $E[\bar{X} | S_i = 0] < E[\bar{X} | S_i = 1]$  and we have that  $t_{SPA,\alpha} < \alpha t_{SPA} + (1-\alpha)\frac{1}{2}$ , as desired.

□

#### 5.8.4 Deriving the optimal Threshold in the Binomial Case

We have for the log-likelihood ratio:

$$\log \frac{P(\bar{S} = s | T = 1)}{P(\bar{S} = s | T = 0)} \equiv LR(\bar{s}) = N \left[ \bar{s} \log \left( \frac{p}{q} \right) - (1-\tau) \log \left( \frac{1-q}{1-p} \right) \right] + \log \left( \frac{\pi}{1-\pi} \right)$$

solving for  $LR(t^*) = 0$ , gives the optimal threshold

$$t^* = \frac{\log \left( \frac{1-q}{1-p} \right)}{\log \left( \frac{1-q}{1-p} \right) + \log \left( \frac{p}{q} \right)} + \frac{\log \left( \frac{\pi}{1-\pi} \right)}{2N \left[ \log \left( \frac{1-q}{1-p} \right) + \log \left( \frac{p}{q} \right) \right]}$$

## 5.9 Task Interface for CONF, PRED and THRESH

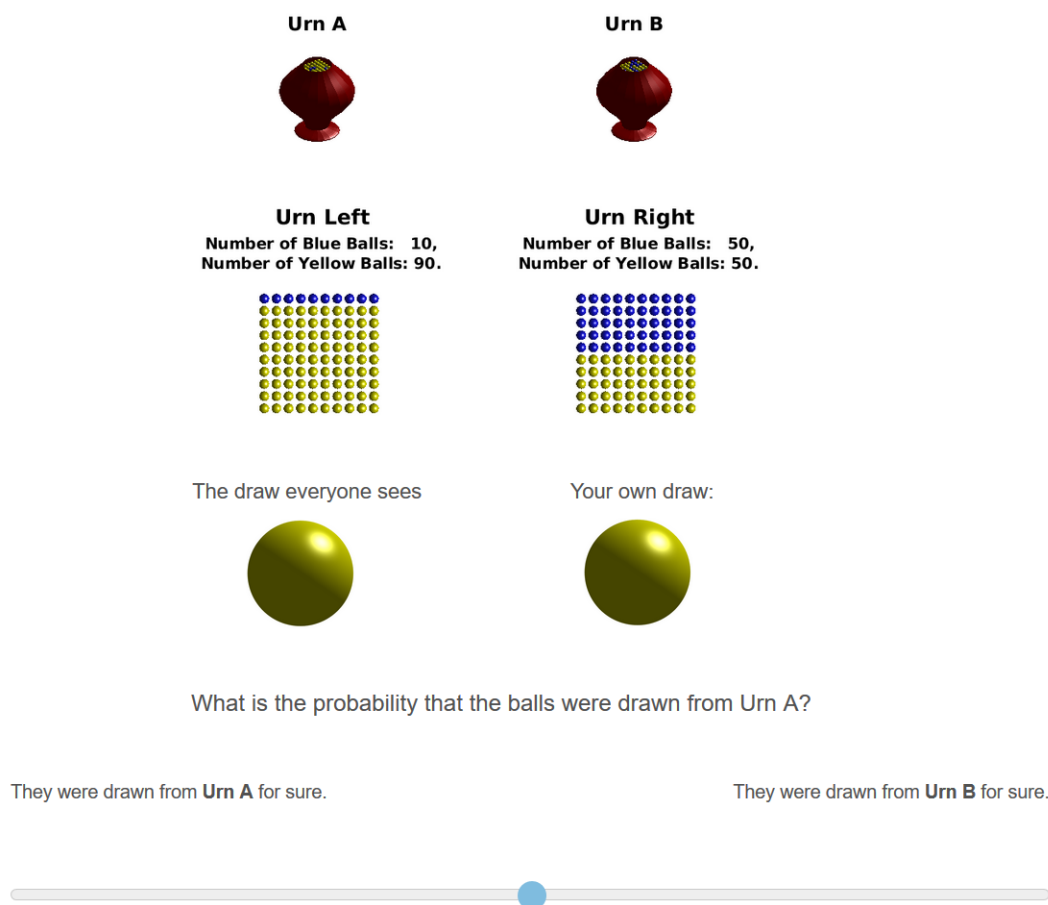




Figure 5.9.1: The interface in treatment CONF for shared information. For providing a confidence level, respondents use a slider which moves from a probability of 100 (Urn A for sure) to a probability of 0 (Urn B for sure).

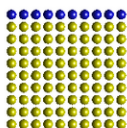
**Urn A**



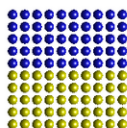
**Urn B**



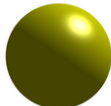
**Urn Left**  
 Number of Blue Balls: 10,  
 Number of Yellow Balls: 90.



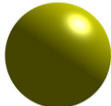
**Urn Right**  
 Number of Blue Balls: 50,  
 Number of Yellow Balls: 50.



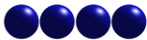
The draw everyone sees

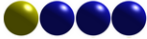


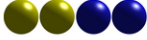
Your own draw:





We also drew a ball randomly for 4 other participants **from the same urn as yours**. How many of these do you expect to be yellow?

0 

1 

2 

3 

4 

☐

☐

☐

☐

☐

Figure 5.9.2: The interface in treatment PRED for shared information. For their prediction, respondents can choose an integer value from 0 to 4.

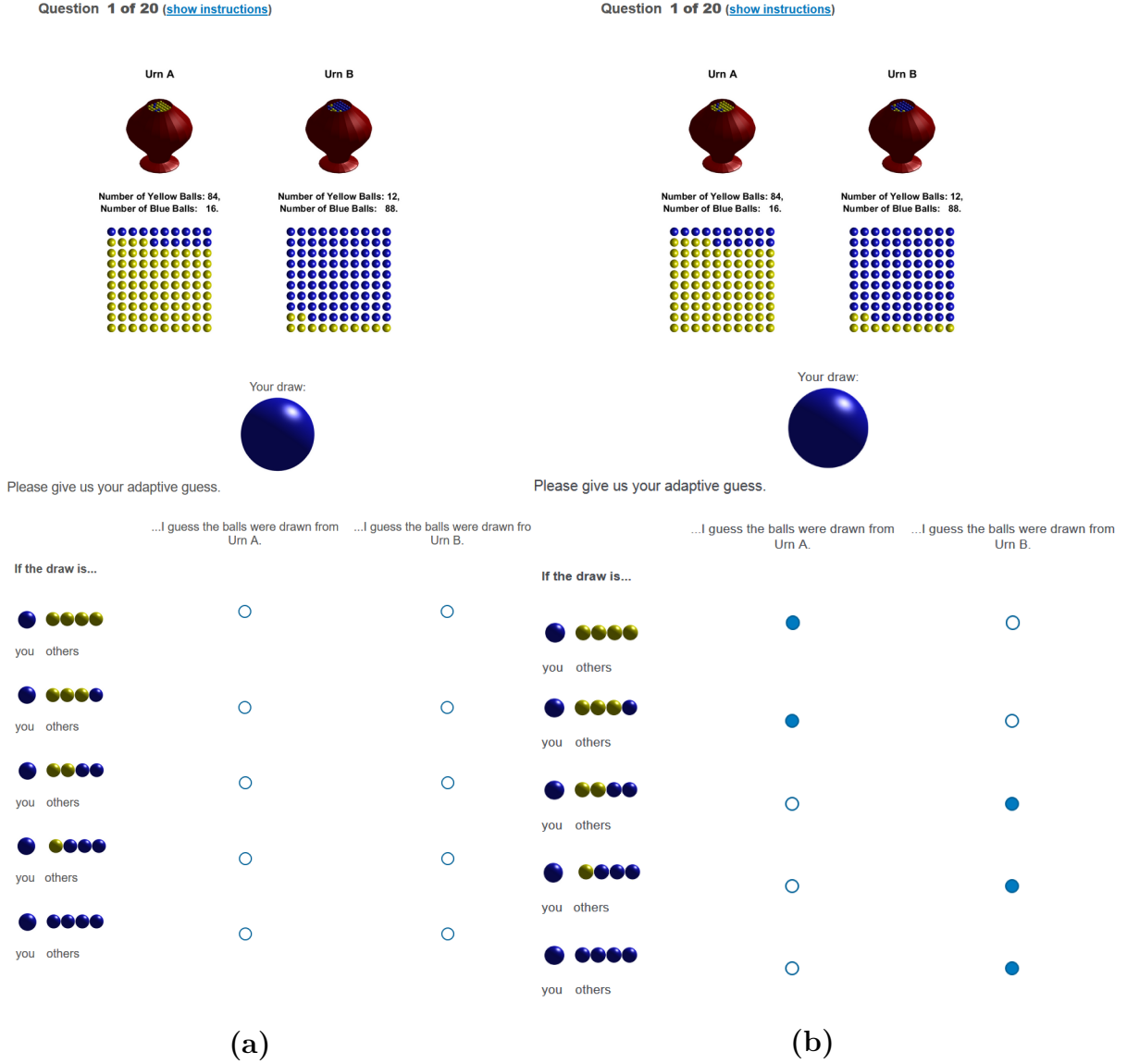


Figure 5.9.3: The interface in treatment THRESH in the private information case, with part (a) showing the initial task a respondent faces and part (b) showing a possible response. When a respondent selects one of the urns for a specific row, the interface automatically selects Urn A for all rows above and Urn B for all rows below. To produce the pattern shown in (b), it is thus enough to select Urn A in the second row or Urn B in the third one. This behavior ensures that responses can always be described by a single threshold level and makes the task more comparable to other treatments in terms of required effort.

The threshold level inferred from the response shown in (b) is 3. If the respondent's own draw was yellow, the inferred threshold level would be 4.

# Chapter 6

## Conclusion

This thesis discussed the properties and performance of various methods targeted at increasing the performance of human crowds. All proposed methods were accompanied by theoretical arguments showing optimality in some sense for a crowd of Bayesian decision-makers. All methods also aimed to be relatively simple, so that they can be used by ‘real’ human beings, and not only by abstract Bayesian decision-makers.

Apart from those similarities, the exact target of each method, and the nature of the results presented in each chapter, were quite different. Chapter 2 showed that the logarithmic market scoring rule (LMSR) proposed by Hanson (2003) is robust to strategic manipulations in a setting in which information is received repeatedly. The LMSR is an already existing method for incentivizing and aggregating forecasts about events which can be verified in the near future. In contrast, Chapter 3 analyzed Baillon’s (2017) Bayesian markets, a method which was proposed for incentivizing forecasts which can *not* be verified in this sense, and related the method to aggregation. Chapter 4 proposed choice-matching, which targets incentives in a more general way. In contrast to the methods in Chapter 1 and 2, choice-matching can be seen as a ‘class’ of methods: we gave numerous examples for applications in areas as distinct as marketing, experimental economics and macroeconomics forecasting.

A new method, self-aggregation (SELF), was also proposed in Chapter 5, which was concerned with the aggregation properties of different ways of combining individual judgments into a single decision. The chapter also has broader implications by demonstrating the success of SELF and Prelec et al.’s (2017) Surprisingly Popular Algorithm (SPA). Both of these ask members of the crowd to answer demanding meta-cognitive questions. The space of such questions seems largely unexplored, with almost the entire previous literature focused on judgments of confidence. Chapter 5 suggests that there may be large performance increases hiding in this unexplored area.

Moreover, while the contributions of Chapter 2 and 4 are theoretical – mostly using tools from game theory – Chapters 3 and 5 also devote much effort to testing the methods in practice. These empirical tests underline the importance of simple methods. While in both experiments, the methods under investigation perform well in the aggregate, we also see several deviations by individual decision-makers from the Bayesian prescription. Developing a systematic framework of these deviations and how they affect the ranking of different crowd methods looks like a promising direction of future theoretical work. Some steps in this direction were made with the anchoring and adjustment model in Chapter 5.

Future applications of crowd-sourcing of the type suggested in the Introduction will need to address all issues discussed in this thesis (and presumably many more). It is therefore natural to think of ways of combining the different methods proposed here into a single application. For instance, in Chapter 4 it is suggested to link incentive schemes for forecasts of verifiable and unverifiable events. Using the space mission example from the Introduction, we could for instance ask a crowd to make forecasts about a successful manned mission to another star system before the year 2200. Simultaneously, there could be a prediction market (which could be of the type analyzed in Chapter 2) for verifiable, short-term developments, such as the question whether there will be another Moon landing, or a successful human Mars mission before the year 2035. For each individual crowd



member, the payoff for their long-term, unverifiable forecasts could then be the average payoff on the short-term prediction markets of those crowd members who made the same long-term prediction as them. One could also try to find a market implementation for the long-term forecasts, such as the ones discussed in Chapter 3. Chapter 5 in turn would suggest that on such a market, it could be beneficial to trade assets which are more complex than whether a certain event occurs or does not occur. For instance – as a market analogy for the meta-cognitive statements discussed in Chapter 5 – one could allow the trade of ‘derivatives’ in the market, such as predictions about the development of future prices, or about the number of participants.

It is hard to think of the intergovernmental decision-making processes which determine whether countries are at war or at peace with each other, how to combat climate change, how to regulate new transformative technologies or how to conduct international trade – and believe that these processes ensure all relevant perspectives are reflected in the eventual decision. I suspect that the coming decades will see many innovations with respect to collective decision-making and that most of these innovations will make use of ever-more-powerful information technology to aggregate the knowledge and skill possessed by millions of people in real-time. Ideally, the chapters in this thesis can contribute to our understanding of what shape these innovations will take.



# Summary

This dissertation deals with various methods of rewarding and collecting judgments from ‘crowds’ (large groups of people). The two main challenges considered throughout this thesis are the ‘incentive problem’ and the ‘aggregation problem’. The incentive problem is the problem of choosing rewards for disclosing and acquiring information. The aggregation problem involves mapping the individual judgments of all crowd members into one decision. In Chapter 2, I analyze, using game-theoretic tools, the strategic aspects of the Logarithmic Market Score Rule (LMSR). I extend existing theoretical arguments in favor of the LMSR, in particular allowing for a dynamic in-flow of information. In addition to the formal arguments, I also informally discuss various other practical aspects of implementing the LMSR.

Unlike Chapter 2, Chapter 3 deals with beliefs about events that are not verifiable, i.e. where there is no way to determine (later) whether an answer was correct or not. Just as the LMSR discussed in Chapter 2 tries to simultaneously solve the incentive problem and the aggregation problem for predictions of verifiable events, the Follow the money (FTM) mechanism studied in Chapter 3 addresses these two problems for non-verifiable events. We test the FTM in a large online sample of US students, using a task that ensures that the informational assumptions of the model are met. The experiment shows evidence that the FTM is successful in providing incentives for truth-telling and in aggregating opinions, although there are also substantial deviations from the theoretical predictions in respondent behavior.

Chapter 4 continues the discussion on incentives for revealing unverifiable beliefs and expands the incentive scheme discussed in Chapter 3, allowing for any number of possible answers. The idea is to link explicit opinions or judgments, which are responses to a multiple choice question (MCQ), with an auxiliary task that reveals the respondents’ “types”, but only implicitly. We focus on a case where the auxiliary task asks respondents to predict how often each answer option in the MCQ was chosen by the other respondents. A respondent’s score is then a weighted sum of a prediction accuracy score and the average prediction accuracy score of all respondents endorsing the same answer to the MCQ. We show that under plausible assumptions about the updating process, truth-telling is a strict Bayes-Nash equilibrium.

Chapter 5 shifts the focus again to the aggregation problem. I propose “self-aggregation” (SELF) which asks individuals to vote for one of two options while setting a threshold for the number of people who would convince them of the other option. SELF then chooses an option if the number of people voting in favor exceeds the average threshold in the group. In a model in which individuals update their beliefs in a (roughly) Bayesian way, I show that SELF is predicted to outperform the benchmarks. In an experimental test of the model, respondents solve a binary decision problem in a stylized urn experiment, in which reactions and aggregation results can be directly compared to the Bayesian prescription. In the experiment, SELF compares favorably with confidence-based methods, but does not realize its theoretical advantage over the so-called surprisingly popular algorithm (SPA). The results show that although individuals’ metacognitive skills are challenged by complex methods such as SELF and the SPA, the answers contain enough information to outperform methods based on less challenging questions.

# Samenvatting

Dit proefschrift behandelt verschillende methoden om oordelen van ‘crowds’ (grote groepen mensen) te belonen en te aggregeren. De twee belangrijkste uitdagingen die in dit proefschrift aan bod komen, zijn het ‘stimuleringsprobleem’ en het ‘aggregatieprobleem’. Het stimuleringsprobleem is het probleem van het kiezen van beloningen voor het vrijgeven en verkrijgen van informatie. Het aggregatieprobleem omvat het in kaart brengen van de individuele oordelen van alle leden van het publiek in één beslissing. In het eerste hoofdstuk analyseer ik met speltheoretische methodes de strategische aspecten van de Logarithmic Market Scoring Rule (LMSR). Ik breid bestaande theoretische argumenten uit ten gunste van de LMSR, in het bijzonder het toelaten van een dynamische informatiestroom. Naast de formele argumenten bespreek ik ook informeel diverse andere praktische aspecten van het implementeren van de LMSR.

In tegenstelling tot hoofdstuk 2, behandelt hoofdstuk 3 meningen over gebeurtenissen die niet verifieerbaar zijn, d.w.z. waar er geen manier is om (later) te bepalen of een antwoord juist was of niet. Net zoals dat de LMSR besproken in hoofdstuk 2 probeert om tegelijkertijd het stimuleringsprobleem en het aggregatieprobleem voor voorspellingen van verifieerbare gebeurtenissen op te lossen, behandelt het Follow the Money (FTM) mechanisme dat in hoofdstuk 3 wordt bestudeerd deze twee problemen voor niet-verifieerbare gebeurtenissen. We testen de FTM in een grote online steekproef van Amerikaanse studenten, met behulp van een taak die ervoor zorgt dat aan de aannames van het model wordt voldaan. Het experiment toont aan dat de FTM erin slaagt om prikkels te geven voor

het vertellen van de waarheid en om meningen te aggregeren, hoewel er ook substantiële afwijkingen zijn van de theoretische voorspellingen in het gedrag van respondenten.

Hoofdstuk 4 vervolgt de discussie over prikkels om niet-verifieerbare overtuigingen aan het licht te brengen en breidt het stimuleringschema uit dat in Hoofdstuk 3 besproken wordt uit, waarbij een groot aantal mogelijke antwoorden mogelijk is. Het idee is om expliciete meningen of oordelen, die antwoorden op een meerkeuzevraag (MCQ) zijn, te koppelen aan een bijkomende taak die de 'typen' van de respondenten onthult, maar alleen impliciet. We richten ons op een casus waarin de hulptaak respondenten vraagt om te voorspellen hoe vaak elke antwoordoptie in de MCQ door de andere respondenten is gekozen. De score van een respondent is dan een gewogen som van een score voor voorspellingsnauwkeurigheid en de gemiddelde score voor voorspellingsnauwkeurigheid van alle respondenten die hetzelfde antwoord op de MCQ geven. We laten zien dat onder plausibele aannames over het updateproces, het vertellen van de waarheid een strikt Bayes-Nash-evenwicht is.

Hoofdstuk 5 verlegt de focus weer naar het aggregatieprobleem. Ik stel 'zelfaggregatie' (SELF) voor, waarbij individuen worden gevraagd om voor één uit twee opties te stemmen, terwijl ze een drempel instellen voor het aantal mensen dat hen van de andere optie zou overtuigen. SELF kiest vervolgens een optie als het aantal mensen dat voor stemt de gemiddelde drempel in de groep overschrijdt. In een model waarin individuen hun overtuigingen op een (ruwweg) Bayesiaanse manier actualiseren, laat ik zien dat SELF in voorspelling beter presteert dan de benchmarks. In een experimentele test van het model lossen respondenten een binair beslissingsprobleem op in een gestileerd urn-experiment, waarin reacties en aggregatieresultaten direct kunnen worden vergeleken met de Bayesiaanse aanbeveling. In het experiment steekt SELF gunstig af bij op vertrouwen gebaseerde methoden, maar realiseert het zijn theoretische voordeel niet ten opzichte van het zogenaamde Surprisingly Popular Algorithm (SPA). De resultaten laten zien dat hoewel de metacognitieve vaardigheden van individuen worden uitgedaagd door complexe metho-

den zoals SELF en de SPA, de antwoorden voldoende informatie bevatten om beter te presteren dan methoden gebaseerd op minder uitdagende vragen.





The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

- 723        N.M. BOSCH, *Empirical Studies on Tax Incentives and Labour Market Behaviour*
- 724        S.D. JAGAU, *Listen to the Sirens: Understanding Psychological Mechanisms with Theory and Experimental Tests*
- 725        S. ALBRECHT, *Empirical Studies in Labour and Migration Economics*
- 726        Y.ZHU, *On the Effects of CEO Compensation*
- 727        S. XIA, *Essays on Markets for CEOs and Financial Analysts*
- 728        I. SAKALAUSKAITE, *Essays on Malpractice in Finance*
- 729        M.M. GARDBERG, *Financial Integration and Global Imbalances.*
- 730        U. THÜMMEL, *Of Machines and Men: Optimal Redistributive Policies under Technological Change*
- 731        B.J.L. KEIJSERS, *Essays in Applied Time Series Analysis*
- 732        G. CIMINELLI, *Essays on Macroeconomic Policies after the Crisis*
- 733        Z.M. LI, *Econometric Analysis of High-frequency Market Microstructure*
- 734        C.M. OOSTERVEEN, *Education Design Matters*
- 735        S.C. BARENDSE, *In and Outside the Tails: Making and Evaluating Forecasts*

- 736 S. SÓVÁGÓ, *Where to Go Next? Essays on the Economics of School Choice*
- 737 M. HENNEQUIN, *Expectations and Bubbles in Asset Market Experiments*
- 738 M.W. ADLER, *The Economics of Roads: Congestion, Public Transit and Accident Management*
- 739 R.J. DÖTTLING, *Essays in Financial Economics*
- 740 E.S. ZWIERS, *About Family and Fate: Childhood Circumstances and Human Capital Formation*
- 741 Y.M. KUTLUAY, *The Value of (Avoiding) Malaria*
- 742 A. BOROWSKA, *Methods for Accurate and Efficient Bayesian Analysis of Time Series*
- 743 B. HU, *The Amazon Business Model, the Platform Economy and Executive Compensation: Three Essays in Search Theory*
- 744 R.C. SPERNA WEILAND, *Essays on Macro-Financial Risks*
- 745 P.M. GOLEC, *Essays in Financial Economics*
- 746 M.N. SOUVERIJN, *Incentives at work*
- 747 M.H. COVENEY, *Modern Imperatives: Essays on Education and Health Policy*
- 748 P. VAN BRUGGEN, *On Measuring Preferences*
- 749 M.H.C. NIENTKER, *On the Stability of Stochastic Dynamic Systems and their use in Econometrics*
- 750 S. GARCIA MANDICÓ, *Social Insurance, Labor Supply and Intra-Household Spillovers*

- 751 Y. SUN, *Consumer Search and Quality*
- 752 I. KERKEMEZOS, *On the Dynamics of (Anti) Competitive Behaviour in the Airline Industry*
- 753 G.W. GOY, *Modern Challenges to Monetary Policy*
- 754 A.C. VAN VLODROP, *Essays on Modeling Time-Varying Parameters*
- 755 J. SUN, *Tell Me How To Vote, Understanding the Role of Media in Modern Elections*
- 756 J.H. THIEL, *Competition, Dynamic Pricing and Advice in Frictional Markets: Theory and Evidence from the Dutch Market for Mortgages*
- 757 A. NEGRIU, *On the Economics of Institutions and Technology: a Computational Approach*
- 758 F. GRESNIGT, *Identifying and Predicting Financial Earth Quakes using Hawkes Processes*
- 759 A. EMIRMAHMUTOGLU, *Misperceptions of Uncertainty and Their Applications to Prevention*
- 760 A. RUSU, *Essays in Public Economics*
- 761 M.A. COTOFAN, *Essays in Applied Microeconomics: Non-Monetary Incentives, Skill Formation, and Work Preferences*
- 762 B.P.J. ANDRÉE, *Theory and Application of Dynamic Spatial Time Series Models*
- 763 P. PELZL, *Macro Questions, Micro Data: The Effects of External Shocks on Firms*

- 764 D.M. KUNST, *Essays on Technological Change, Skill Premia and Development*
- 765 A.J. HUMMEL, *Tax Policy in Imperfect Labor Markets*
- 766 T. KLEIN, *Essays in Competition Economics*
- 767 M. VIGH, *Climbing the Socioeconomic Ladder: Essays on Sanitation and Schooling*
- 768 YAN XU, *Eliciting Preferences and Private Information: Tell Me What You Like and What You Think*
- 769 S. RELLSTAB, *Balancing Paid Work and Unpaid Care over the Life-Cycle*
- 770 Z. DENG, *Empirical Studies in Health and Development Economics*
- 771 L. KONG, *Identification Robust Testing in Linear Factor Models*
- 772 I. NEAMȚU, *Unintended Consequences of Post-Crisis Banking Reforms*
- 773 B. KLEIN TEESELINK, *From Mice to Men: Field Studies in Behavioral Economics*

# Bibliography

- Agarwal, A., Mandal, D., Parkes, D. C., and Shah, N. (2020). Peer prediction with heterogeneous users. *ACM Transactions on Economics and Computation (TEAC)*, 8(1):1–34.
- Anderson, L. R. and Holt, C. A. (1997). Information cascades in the laboratory. *The American Economic Review*, 87(5):847–862.
- Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., Neumann, G. R., Ottaviani, M., Schelling, T. C., Shiller, R. J., Smith, V. L., Snowberg, E., Sunstein, C. R., Tetlock, P. C., Tetlock, P. E., Varian, H. R., Wolfers, J., and Zitzewitz, E. (2008). The promise of prediction markets. *Science*, 320(5878):877–878.
- Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, pages 1236–1239.
- Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, 114(30):7958–7962.
- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145.
- Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.
- Berg, J., Forsythe, R., Nelson, F., and Rietz, T. (2008). Results from a dozen years of

- election futures markets research. *Handbook of experimental economics results*, 1:742–751.
- Black, D. (1948). On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34.
- Buckley, P. and O’Brien, F. (2017). The effect of malicious manipulations on prediction market accuracy. *Information Systems Frontiers*, 19(3):611–623.
- Budescu, D. V. and Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637.
- Carvalho, A. (2017). On a participation structure that ensures representative prices in prediction markets. *Decision Support Systems*, 104:13–25.
- Chambers, C. P. and Lambert, N. S. (2020). Dynamic belief elicitation. *Econometrica*. forthcoming.
- Chen, Y., Dimitrov, S., Sami, R., Reeves, D. M., Pennock, D. M., Hanson, R. D., Fortnow, L., and Gonen, R. (2010). Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, 58(4):930–969.
- Chen, Y. and Pennock, D. M. (2007). A utility framework for bounded-loss market makers. *arXiv preprint arXiv:1206.5252*.

- Chen, Y. and Pennock, D. M. (2010). Designing markets for prediction. *AI Magazine*, 31(4):42–52.
- Chen, Y., Reeves, D. M., Pennock, D. M., Hanson, R. D., Fortnow, L., and Gonen, R. (2007). Bluffing and strategic reticence in prediction markets. In *International Workshop on Web and Internet Economics*, pages 70–81. Springer.
- Chen, Y. and Vaughan, J. W. (2010). A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198.
- Compte, O. and Postlewaite, A. (2018). *Ignorance and Uncertainty*. Cambridge University Press.
- Condorcet, M. d. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. de l’Imprimerie Royale.
- Cowgill, B., Wolfers, J., and Zitzewitz, E. (2009). Using prediction markets to track information flows: Evidence from google. In *AMMA*, page 3.
- Cowgill, B. and Zitzewitz, E. (2015). Corporate prediction markets: Evidence from google, ford, and firm x. *The Review of Economic Studies*, 82(4):1309–1341.
- Cvitanic, J., Prelec, D., Radas, S., and Sikic, H. (2020). Incentive-compatible surveys via posterior probabilities. *Theory of Probability & Its Applications*, 65(2):292–321.
- Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25(1):1–17.
- de Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:298–329. Translated into English by Mara Khale as “On the Subjective Meaning of Probability”, reprinted in Monari and Cocchi (1993), pp. 291–321.

- DellaVigna, S. and Pope, D. (2017). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.
- DellaVigna, S. and Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, 126(6):2410–2456.
- Dimitrov, S. and Sami, R. (2008). Non-myopic strategies in prediction markets. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 200–209.
- Dimitrov, S., Sami, R., and Epelman, M. A. (2015). Subsidized prediction mechanisms for risk-averse agents. *ACM Transactions on Economics and Computation (TEAC)*, 3(4):1–23.
- Douceur, J. R. (2002). The sybil attack. In *International workshop on peer-to-peer systems*, pages 251–260. Springer.
- Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347.
- Dryzek, J. S. and List, C. (2003). Social choice theory and deliberative democracy: a reconciliation. *British Journal of Political Science*, 33(1):1–28.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.
- Fan, Y., Budescu, D. V., Mandel, D., and Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis*, 16(3):197–217.



- Feddersen, T. and Pesendorfer, W. (1998). Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review*, 92(1):23–35.
- Forsythe, R., Nelson, F., Neumann, G. R., and Wright, J. (1992). Anatomy of an experimental political stock market. *The American Economic Review*, 82(5):1142–1161.
- Frank, M. R., Cebrian, M., Pickard, G., and Rahwan, I. (2017). Validating bayesian truth serum in large-scale online human experiments. *PloS one*, 12(5):e0177385.
- Friedman, D. (1983). Effective scoring rules for probabilistic forecasts. *Management Science*, 29(4):447–454.
- Galton, F. (1907). Vox populi. *Nature*, 75(7):450–451.
- Genest, C., Zidek, J. V., et al. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135.
- Gillen, B. J., Plott, C. R., and Shum, M. (2017). A pari-mutuel-like mechanism for information aggregation: A field test inside intel. *Journal of Political Economy*, 125(4):1075–1099.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
- Gneiting, T., Ranjan, R., et al. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.
- Hankins, R. and Lee, A. (2011). Crowd sourcing and prediction markets. Working paper.
- Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119.

- Hanson, R. (2007). Logarithmic markets coring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1):3–15.
- Hanson, R. (2013). Shall we vote on values, but bet on beliefs? *Journal of Political Philosophy*, 21(2):151–178.
- Harsanyi, J. C. (1967–1968). Games with incomplete information played by bayesian players. *Management Science*, 14:159–189.
- Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, 35(4):519–530.
- Hertwig, R. (2012). Tapping into the wisdom of the crowds – with confidence. *Science*, 336(6079):303–304.
- Herzog, S. M. and Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2):231–237.
- Herzog, S. M. and Hertwig, R. (2011). The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making*, 6(1):58–72.
- Hurwicz, L. (1960). Decentralization and computation in resource allocation. In Arrow, K. J., editor, *Mathematical methods in the social sciences*. Stanford University Press.
- John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532.
- Johnson, D. D. and Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477(7364):317.

- Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4):237.
- Kahneman, D. and Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, 11(2):123–141.
- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273.
- Kothiyal, A., Spinu, V., and Wakker, P. P. (2010). Comonotonic proper scoring rules to measure ambiguity and subjective beliefs. *Journal of Multi-Criteria Decision Analysis*, 17(3-4):101–113.
- Kyburg, H. E. and Smokler, H. E. (1980). *Studies in subjective probability, 2nd edition*. Robert E. Krieger Publishing. Huntington, NY, USA.
- Lambert, N. (2018). Probability elicitation for agents with arbitrary risk preferences. Technical report, Working Paper.
- Larrick, R. P. and Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, 52(1):111–127.
- Lee, M. D., Danileiko, I., and Vi, J. (2018). Testing the ability of the surprisingly popular method to predict nfl games. *Judgment and Decision Making*, 13(4):322.
- Lichtenstein, S., Fischhoff, B., and Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 306–334. Cambridge University Press, Cambridge, UK.
- List, C. and Goodin, R. E. (2001). Epistemic democracy: Generalizing the condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306.

## BIBLIOGRAPHY

---

- Liu, Y. and Chen, Y. (2017). Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 63–80.
- Macmillan, N. (2002). Signal detection theory. In Pashler, H., editor, *Stevens' Handbook of Experimental Psychology, Methodology in Experimental Psychology*, volume 4. John Wiley and Sons.
- Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters*, 91(3):425–429.
- Marks, G. and Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological bulletin*, 102(1):72.
- Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic theory*, volume 1. Oxford university press New York.
- McCoy, J. and Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint arXiv:1703.04778*.
- Milgrom, P. and Stokey, N. (1982). Information, trade and common knowledge. *Journal of Economic Theory*, 26(1):17–27.
- Miller, N., Resnick, P., and Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373.
- Monari, P. and Cocchi, D. (1993). *Probabilità e Induzione*. Clueb, Bologna.
- Nau, R. F. (1985). Should scoring rules be effective? *Management Science*, 31(5):527–535.
- Offerman, T., Sonnemans, J., Van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, 76(4):1461–1489.

- Okoli, C. and Pawlowski, S. D. (2004). The delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1):15–29.
- Ortner, G. (1998). Forecasting markets: An industrial application. Technical report, mimeo.
- Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.
- Ostrovsky, M. (2012). Information aggregation in dynamic markets with strategic traders. *Econometrica*, 80(6):2595–2647.
- Plott, C. R. and Chen, K.-Y. (2002). Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem.
- Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge.
- Prelec, D. (2004). A bayesian truth serum for subjective data. *Science*, 306(5695):462–466.
- Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535.
- Radanovic, G. and Faltings, B. (2013). A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI’13)*, number CONF, pages 833–839.
- Radanovic, G. and Faltings, B. (2014). Incentives for truthful information elicitation of continuous signals. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI’14)*, number CONF, pages 770–776.
- Radanovic, G. and Faltings, B. (2015). Incentive schemes for participatory sensing. In *Proceedings of the 14th international conference on autonomous agents and multiagent systems (AAMAS’15)*, number CONF, pages 1081–1089.

- Radanovic, G., Faltings, B., and Jurca, R. (2016). Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):1–28.
- Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.
- Ramsey, F. P. (1931). Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–198. Routledge and Kegan Paul, London. Reprinted in Kyburg and Smokler (1980), pp. 61-92.
- Richardson, H. R. and Stone, L. D. (1971). Operations analysis during the underwater search for scorpion. *Naval Research Logistics Quarterly*, 18(2):141–157.
- Ross, L., Greene, D., and House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.
- Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378.
- Roth, A. E. (2018). Marketplaces, markets, and market design. *American Economic Review*, 108(7):1609–58.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.
- Savage, L. J. (1954). The foundations of statistics. *Dover Press, New York*.
- Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

- Shnayder, V., Agarwal, A., Frongillo, R., and Parkes, D. C. (2016). Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196.
- Steyvers, M. and Miller, B. (2015). Cognition and collective intelligence. *Handbook of Collective Intelligence*, page 119.
- Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, pages 1339–1342.
- Sunstein, C. R. (2006). *Infotopia: How many minds produce knowledge*. Oxford University Press.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Swank, J., Swank, O. H., and Visser, B. (2008). How committees of experts interact with the outside world: some theory, and evidence from the fomc. *Journal of the European Economic Association*, 6(2-3):478–486.
- Treynor, J. L. (1987). Market efficiency and the bean jar experiment. *Financial Analysts Journal*, 43(3):50–53.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.
- Visser, B. and Swank, O. H. (2007). On committees of experts. *The Quarterly Journal of Economics*, 122(1):337–372.
- Weaver, R. and Prelec, D. (2013). Creating truth-telling incentives with the bayesian truth serum. *Journal of Marketing Research*, 50(3):289–302.
- Witkowski, J. and Parkes, D. C. (2012a). Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 964–981.

- Witkowski, J. and Parkes, D. C. (2012b). A robust bayesian truth serum for small populations.
- Wolfers, J. and Zitzewitz, E. (2006). Interpreting prediction market prices as probabilities. Technical report, National Bureau of Economic Research.
- Wolfers, J., Zitzewitz, E., and Snowberg, E. (2012). Prediction markets for economic forecasting.
- Zhang, P. and Chen, Y. (2014). Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 245–252.