Advancements in the Elicitation and Aggregation of Private Information

by

Arthur Carvalho

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

There are many situations where one might be interested in eliciting and aggregating the private information of a group of agents. For example, a recommendation system might suggest recommendations based on the aggregate opinions of a group of like-minded agents, or a decision maker might take a decision based on the aggregate forecasts from a group of experts.

When agents are self-interested, they are not necessarily honest when reporting their private information. For example, agents who have a reputation to protect might tend to produce forecasts near the most likely group consensus, whereas agents who have a reputation to build might tend to overstate the probabilities of outcomes they feel will be understated in a possible consensus. Therefore, economic incentives are necessary to incentivize self-interested agents to honestly report their private information.

Our first contribution in this thesis is a scoring method to induce honest reporting of an answer to a multiple-choice question. We formally show that, in the presence of social projection, one can induce honest reporting in this setting by comparing reported answers and rewarding agreements. Our experimental results show that the act of encouraging honest reporting through the proposed scoring method results in more accurate answers than when agents have no direct incentives for expressing their true answers.

Our second contribution is about how to incentivize honest reporting when private information are subjective probabilities (beliefs). Proper scoring rules are traditional scoring methods that incentivize honest reporting of subjective probabilities, where the expected score received by an agent is maximized when that agent reports his true belief. An implicit assumption behind proper scoring rules is that agents are risk neutral. In an experiment involving proper scoring rules, we find that human beings fail to be risk neutral. We then start our discussion on how to adapt proper scoring rules to cumulative prospect theory, a modern theory of choice under uncertainty. We explain why a property called comonotonicity is a sufficient condition for proper scoring rules to be indeed proper under cumulative prospect theory. Moreover, we show how to construct a comonotonic proper scoring rule from any traditional proper scoring rule. We also propose a new approach that uses non-deterministic payments based on proper scoring rules to elicit an agent's true belief when the components that drive the agent's attitude towards uncertainty are unknown.

After agents report their private information, there is still the question on how to aggregate the reported information. Our third contribution in this thesis is an empirical study on the influence of the number of agents on the quality of the aggregate information in

a crowdsourcing setting. We find that both the expected error in the aggregate information as well as the risk of a poor combination of agents decrease as the number of agents increases. Moreover, we find that the top-performing agents are consistent across multiple tasks, whereas worst-performing agents tend to be inconsistent.

Our final contribution in this thesis is a pooling method to aggregate reported beliefs. Intuitively, the proposed method works as if the agents were continuously updating their beliefs in order to accommodate the expertise of others. Each updated belief takes the form of a linear opinion pool, where the weight that an agent assigns to a peer's belief is inversely related to the distance between their beliefs. In other words, agents are assumed to prefer beliefs that are close to their own beliefs. We prove that such an updating process leads to consensus, *i.e.*, the agents all converge towards the same belief. Further, we show that if risk-neutral agents are rewarded using the quadratic scoring rule, then the assumption that they prefer beliefs that are close to their own beliefs follows naturally. We empirically demonstrate the effectiveness of the proposed method using real-world data. In particular, the results of our experiment show that the proposed method outperforms the traditional unweighted average approach and another distance-based method when measured in terms of both overall accuracy and absolute error.

Acknowledgements

I would like to thank my supervisors, Dr. Stanko Dimitrov and Dr. Kate Larson, for their support and patience throughout the past 4 years. I am grateful for the amazing feedback received from the committee members: Dr. Craig Boutilier, Dr. Selcuk Onay, Dr. Daniel Lizotte, and Dr. Pascal Poupart.

I would like to thank Dr. Carol Acton, Dr. Katherine Acheson, Dr. Stefan Rehm, Susan Gow, and Dr. Veronica Austen for providing gold-standard answers for my content-analysis experiment.

I would also like to thank various colleagues for the fruitful discussions that contributed to the composition of this work: Hadi Hosseini, Ashish Hota, Alan Tsang, Shehroz Khan, Simina Branzei, Greg Hines, and Xi Alice Gao.

Dedication

To my lovely wife, Marina, and family, Ulissiano, Remedios, and Jardel.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

There are many scenarios where it is desirable to elicit the private information of an agent. One particular context is asking an expert for his belief regarding the occurrence of future outcomes during a decision-making process. For example, companies rely on predictions about consumer demand and material supply to make their production plans, while weather forecasts provide guidelines for long range or seasonal agricultural planning so that farmers can select crops that are best suited to the anticipated climatic conditions.

Another domain highly dependent on the elicitation of agents' private information is crowdsourcing, which consists of the practice of obtaining relevant information or services from a large group of people. Recent technological advances have facilitated the outsourcing of a variety of tasks to "the crowd", *e.g.*, the development and testing of large software applications, the design of websites, professional translation of documents, *etc.*

Unfortunately, self-interested agents are not necessarily honest when reporting their private information. For example, Nakazono [57] reported that governors of the Federal Open Market Committee tend to report forecasts close to the previous consensus, whereas non-governors tend to report forecasts far away from the previous consensus. Nakazono concluded that both governors and non-governors behave strategically. Agents might also bias their reports towards what they believe is socially desirable, or in a way that makes them look more favorable to the requester. This phenomenon is also known as the social desirability bias. It has been reported that the social desirability bias might account for as much as 10%-70% of the variance in agents' reports [59].

Clearly, a method to promote honest reporting is necessary when agents are self-interested or biased. Henceforth, we refer to such methods as *incentive methods*. Incentive methods use economic incentives (not necessarily money) for promoting honest reporting.

Some incentive methods, however, rely on assumptions that are not necessarily realistic, *e.g.*, that there exists an actual outcome that will be observed at a certain point in the future, and that agents are expected-value maximizers. *An important question in this domain is how to circumvent these limitations and design incentive methods under realistic assumptions.*

Ideally, a requester should elicit and aggregate the private information from many different agents so that biases of individual judgments can offset each other, which may result in more accurate aggregate information. This is the core idea behind collective intelligence, or the "wisdom of crowds", *i.e.*, the aggregate information is, at least in expectation, more accurate than the information of any single individual [81]. Agents are not always in agreement when reporting their private information. Consequently, aggregation methods must be used to combine agents' reports into a single representative report. The aggregation aspect brings two new questions to the elicitation process: (1) *how does the number of agents influence the aggregate report?* and (2) *how to aggregate agents' reports in a meaningful way?*

1.1 Contributions

This thesis addresses each one of the aforementioned questions. Regarding incentive methods, our initial contribution is a scoring method based on proper scoring rules to induce honest reporting of private information without relying on the assumption of future observable outcomes. Our proposed method compares agents' reports and rewards agreement. We formally show that this simple payment structure induces honest reporting when there exists social projection, *i.e.*, when each agent believes that his private information is the most common amongst all other agents.

Our second contribution is an approach to incentivize honest reporting of subjective probabilities (beliefs) without assuming that agents are expected-value maximizers. In particular, we adapt proper scoring rules to cumulative prospect theory (CPT), a modern theory of choice under uncertainty. Besides providing sufficient conditions for proper scoring rules to induce honest reporting under cumulative prospect theory, we also propose a new incentive method that uses non-deterministic payments to elicit an agent's belief when the CPT components that drive the agent's attitude towards uncertainty are unknown.

Regarding the aggregation of agents' reports, our first contribution is a study on the influence of the number of agents on the quality of aggregate reports in a crowdsourcing setting. We find that both the expected error in the aggregate report as well as the risk of

a poor combination of agents decrease as the number of agents increases. Moreover, our results show that restricting the population of agents to 40 or fewer percent top-performing agents is likely to produce more accurate aggregate reports, whereas removing up to 40% of the worst-performing agents can actually make the aggregate reports less accurate. We find that this result holds due to top-performing agents being consistent across multiple tasks, whereas worst-performing agents tend to be inconsistent.

Our final contribution is a method to aggregate agents' reported beliefs. We start by assuming that agents prefer beliefs that are close to their own beliefs. Given this assumption, we propose an algorithm to find a consensual belief. The proposed algorithm works by simulating interactions between agents. In each round, agents disclose their beliefs to others, and they update their own beliefs after observing the disclosed beliefs. We formally show that this updating process leads to a consensual belief. Moreover, we show that if risk-neutral agents are rewarded using a particular proper scoring rule called the quadratic scoring rule, then the assumption that agents prefer beliefs that are close to their own beliefs follows naturally.

It is important to note that our research is focused on settings where agents are not concerned about how their private information will be used in the future, as opposed to situations where externalities exist, such as when agents want to influence the potential decision of a decision maker (*e.g.*,[10, 19, 20]).

1.2 Thesis Outline

Besides this introductory section, this thesis has six more chapters:

- In Chapter 2, we introduce *proper scoring rules*, a class of functions which is the core of our proposed methods to elicit agents' private information, and *Amazon's Mechanical Turk*, the crowdsourcing platform utilized for collecting some data sets used throughout this thesis;
- In Chapter 3, we describe our method for inducing honest reporting of private information without relying on future observable outcomes;
- In Chapter 4, we adapt proper scoring rules to cumulative prospect theory in order to elicit agents' beliefs under a more realistic theory of choice under uncertainty;
- In Chapter 5, we present our empirical study on the influence of the number of agents on the quality of the aggregate reports;

- In Chapter 6, we describe our aggregation method for finding a consensual belief;
- In Chapter 7, we conclude by presenting a summary of the contributions of this thesis.

Chapter 2

Background

In this chapter, we introduce *proper scoring rules*, which is the core of our proposed methods for inducing honest reporting, and *Amazon's Mechanical Turk*, which is the platform utilized for collecting some data sets used throughout this thesis.

2.1 Proper Scoring Rules

Consider a set of exhaustive and mutually exclusive outcomes $\{\theta_1, \ldots, \theta_n\}$, for $n \geq 2$, and a probability vector $\mathbf{q} = (q_1, \ldots, q_n)$, where q_x is the probability value associated with the occurrence of outcome θ_x. A *scoring rule* $R(\mathbf{q}, \theta_x)$ is a function that provides a score for the assessment $\mathbf{q}$ upon observing the outcome θ_x, for $x \in \{1, \ldots, n\}$. A scoring rule is called *strictly proper* when an agent receives his maximum expected score if and only if his stated assessment $\mathbf{q}$ corresponds to his true assessment $\mathbf{p} = (p_1, \ldots, p_n)$ [94]. When $\mathbf{q} = \mathbf{p}$, we say that the agent is honestly reporting his belief. The *expected score* of $\mathbf{q}$ at $\mathbf{p}$ for a real-valued scoring rule $R(\mathbf{q}, \theta_x)$ is:

$$\mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}, \cdot)\right] = \sum_{x=1}^{n} p_x \, R(\mathbf{q}, \theta_x) \tag{2.1}$$

Proper scoring rules have been used directly and indirectly to promote honest reporting in a variety of domains, *e.g.*, when grading students' exams [9], when sharing rewards amongst a set of agents based on peer evaluations [15, 16, 17], when incentivizing agents to accurately estimate their own efforts to accomplish a task [5], to elicit predictions in

patient management and clinical trials [78], to elicit opinions from policy makers regarding the occurrence of political and economic events [83], in financial markets set to aggregate agents' subjective probabilities [35], *etc.* Some of the best known strictly proper scoring rules, together with their scoring ranges, are:

$$\begin{aligned} \text{logarithmic: } & R(\mathbf{q}, \theta_x) = \log q_x && (-\infty, 0] \\ \text{quadratic: } & R(\mathbf{q}, \theta_x) = 2q_x - \sum_{k=1}^{n} q_k^2 && [-1, 1] \\ \text{spherical: } & R(\mathbf{q}, \theta_x) = \frac{q_x}{\sqrt{\sum_{k=1}^{n} q_k^2}} && [0, 1] \end{aligned} \tag{2.2}$$

Selten [74] and Jose [41] provided axiomatic characterizations of, respectively, the quadratic scoring rule and the spherical scoring rule in terms of desirable properties, *e.g.*, sensitivity to small differences of small probabilities, symmetry, *etc.*

For the sake of illustration, consider a coin toss experiment with two outcomes ($n = 2$): θ_1 = "heads" and θ_2 = "tails". Consider that an agent i has a true assessment $\mathbf{p} = (0.4, 0.6)$. Assume that agent i reports the assessment $\mathbf{q} = (q_1, q_2)$, which is rewarded according to the logarithmic scoring rule. Then, agent i's expected score is $\mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}, \cdot)\right] = p_1 \log q_1 + p_2 \log q_2 = 0.4 \log q_1 + 0.6 \log q_2$. In the future, if outcome θ_1 is the observed outcome, then the score agent i receives is equal to $\log q_1$. Since the logarithmic scoring rule is a strictly proper scoring rule, agent i's expected score is strictly maximized when he is honest, *i.e.*, when $\mathbf{q} = \mathbf{p} = (0.4, 0.6)$. To show this, note that $\mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}, \cdot)\right] = p_1 \log q_1 + p_2 \log q_2 = p_1 \log q_1 + (1 - p_1) \log(1 - q_1)$. Since the resulting function is concave, its maximum value can be found by taking the first-order derivative with respect to q_1, and equating the result to zero, *i.e.*:

$$\frac{p_1}{q_1} - \frac{1 - p_1}{1 - q_1} = 0 \implies p_1 = q_1$$

The above proper scoring rules are all *symmetric*, in a sense that $R((q_1, \ldots, q_n), \theta_x) = R((q_{\pi_1}, \ldots, q_{\pi_n}), \theta_{\pi_x})$, for all probability vectors $\mathbf{q} = (q_1, \ldots, q_n)$, for all permutations π on n elements, and for all outcomes indexed by $x \in \{1, \ldots, n\}$. We say that a scoring rule is *bounded* if $R(\mathbf{q}, \theta_x) \in \mathbb{R}$, for all probability vectors $\mathbf{q}$ and $x \in \{1, \ldots, n\}$. For example, the logarithmic scoring rule is not bounded because it might return $\log 0 = -\infty$ whenever the probability vector $\mathbf{q}$ contains a probability value equal to zero. On the other hand, both the quadratic and the spherical scoring rules are bounded.

We say that a scoring rule is *positive* when all the returned scores are nonnegative, *i.e.*, $R(\mathbf{q}, \theta_x) \geq 0$ for all $x \in \{1, \ldots, n\}$. A *negative* scoring rule only returns nonpositive scores, *i.e.*, $R(\mathbf{q}, \theta_x) \leq 0$ for all $x \in \{1, \ldots, n\}$. Finally, a *mixed* scoring rule might return both positive and negative scores.

A well-known property of strictly proper scoring rules is that they are still strictly proper under positive affine transformations, *i.e.*, $\gamma R(\mathbf{q}, \theta_x) + \lambda$, for $\gamma > 0$ and $\lambda \in \mathbb{R}$, is also strictly proper [34]. We prove this statement below for the sake of completeness.

Proposition 1. *If $R(\mathbf{q}, \theta_x)$ is a strictly proper scoring rule, then a positive affine transformation of R, i.e., $\gamma R(\mathbf{q}, \theta_x) + \lambda$, for $\gamma > 0$ and $\lambda \in \Re$, is also strictly proper.*

Proof. We just need to note that $\arg\max_{\mathbf{q}} \mathbb{E}\left[\gamma R(\mathbf{q}, \cdot) + \lambda\right] = \arg\max_{\mathbf{q}} \mathbb{E}\left[R(\mathbf{q}, \cdot)\right]$, which is equal to the true assessment $\mathbf{p}$ since R is a strictly proper scoring rule. □

In a seminal work, Savage [72] showed that any differentiable strictly convex function $J(\mathbf{q})$ that is well-behaved at the endpoints of the scoring range can be used to generate a proper scoring rule. Formally:

$$R(\mathbf{q}, \theta_x) = J(\mathbf{q}) - \left(\sum_k \frac{\partial J(\mathbf{q})}{\partial q_k} \times q_k \right) + \frac{\partial J(\mathbf{q})}{\partial q_x}$$

For example, consider $J(\mathbf{q}) = \sum_k q_k \log q_k$. The resulting proper scoring rule is then the logarithmic scoring rule:

$$\begin{aligned} R(\mathbf{q}, \theta_x) &= \sum_k q_k \log q_k - \left(\sum_k (\log q_k + 1) \times q_k \right) + \log q_x + 1 \\ &= \log q_x \end{aligned}$$

Gneiting and Raftery [34] generalize Savage's results in terms of subgradient methods. Scoring rules are also classified based on monotonicity properties. Consider a metric G that assigns to any pair of assessments $\mathbf{q}_1$ and $\mathbf{q}_2$ a real number, which in turn can be seen as the shortest distance between $\mathbf{q}_1$ and $\mathbf{q}_2$. By definition, a metric G must satisfy the following conditions for any assessments $\mathbf{q}_i, \mathbf{q}_j$, and $\mathbf{q}_k$:

1. Positivity: $G(\mathbf{q}_i, \mathbf{q}_j) \geq 0$, for all i, j, and $G(\mathbf{q}_i, \mathbf{q}_j) = 0$ if and only if $\mathbf{q}_i = \mathbf{q}_j$;

2. Symmetry: $G(\mathbf{q}_i, \mathbf{q}_j) = G(\mathbf{q}_j, \mathbf{q}_i)$;

3. Triangle Inequality: $G(\mathbf{q}_i, \mathbf{q}_k) \leq G(\mathbf{q}_i, \mathbf{q}_j) + G(\mathbf{q}_j, \mathbf{q}_k)$.

We say that a scoring rule R is *effective* with respect to G if the following relation holds for any assessments $\mathbf{q}_1$ and $\mathbf{q}_2$ [30]:

$$G(\mathbf{q}_1, \mathbf{p}) < G(\mathbf{q}_2, \mathbf{p}) \iff \mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}_1, \cdot)\right] > \mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}_2, \cdot)\right]$$

In words, each agent's expected score can be seen as a monotone decreasing function of the distance between his true assessment and the reported one, *i.e.*, agents still strictly maximize their expected scores by reporting honestly, and the closer a reported assessment is to the true assessment according to the metric G, the higher the expected score will be. Consequently, the property of effectiveness is stronger than strict properness, and it has been proposed as a desideratum for scoring rules for reasons of monotonicity in keeping an agent close to his true assessment. Friedman [30] discussed some examples of effective scoring rules. For example, the quadratic scoring rule is effective with respect to the root-mean-square deviation, the spherical scoring rule is effective with respect to a renormalized L_2-metric, and the logarithmic scoring rule is not effective with respect to any metric [58].

2.2 Amazon's Mechanical Turk

Amazon's Mechanical Turk[1] (AMT) is an online labor market originally developed for human computation tasks, *i.e.*, tasks that are relatively easy for human beings, but nonetheless challenging or even currently impossible for computers, *e.g.*, audio transcription, filtering adult content, extracting data from images, *etc.* Some studies have shown that AMT can effectively collect valid data in these settings [50, 77].

The standard crowdsourcing process on AMT works as follows: first, a requester asks a number of workers to individually complete a common task. After completing the task, each worker must report back an output. Finally, each worker receives from the requester a common baseline payment and, sometimes, an individual bonus.

Recently, AMT has also been used as a platform for conducting behavioral experiments [52]. One of the advantages that it offers to researchers is the access to a large, diverse, and stable pool of people willing to participate in the experiments for relatively low pay,

[1] https://www.mturk.com

thus simplifying the recruitment process and allowing a faster iteration between developing theory and executing experiments. The majority of workers on AMT come from the United States, but with an increasing proportion coming from India [40].

AMT also provides an easy-to-use built-in mechanism to pay workers that greatly reduces the difficulties of compensating individuals for their participation in the experiments, and a built-in reputation system that helps requesters distinguish between good and bad workers and, consequently, to ensure data quality. Many studies have shown that results of behavioral studies conducted on AMT are comparable to results obtained in other online domains as well as in offline settings [12, 38], thus providing evidence that AMT is a valid means of collecting behavioral data.

As concrete examples of how AMT has been used in practice, we can cite the social network LinkedIn, which has been using workers from AMT for transcribing business cards[2], the Defense Advanced Research Projects Agency (DARPA) uses AMT to translate Arabic texts to English[3], the mobile-web community SnapMyLife has been using AMT to tag photos[4], *etc.*

[2] https://requester.mturk.com/case_studies/cs/linkedin
[3] https://requester.mturk.com/case_studies/cs/darpa
[4] https://requester.mturk.com/case_studies/cs/snapmylife

Chapter 3

Inducing Honest Reporting Without a Ground Truth

Traditional scoring methods that incentivize honest reporting, such as proper scoring rules, rely on the assumption that there is an observable future outcome, or a ground truth. Such an assumption is not always reasonable. For example, when market analysts provide sales forecasts on a potential new product, there is no guarantee that the product will ever be produced. Hence, the actual number of sales may never be observed. Crowdsourcing is another domain where the assumption of the existence of a ground-truth outcome is not always reasonable. After eliciting information from the crowd, verifying the accuracy of each report can be either too costly or not even possible because the underlying task might be subjective, such as in content and sentiment analysis.

In this chapter, we propose a new scoring method based on proper scoring rules for inducing honest reporting when agents have to answer a multiple-choice question. However, different than what is traditionally assumed in the proper scoring rules literature, the proposed method does not assume that there is a ground-truth answer. Instead, scores are determined based on pairwise comparisons between agents' reported answers. Our scoring method exploits the psychological concept of *social projection* to induce honest reporting, *i.e.*, the tendency of agents to expect similarities between themselves and others.

The setting, as we consider in this chapter, can be described as follows: a requester elicits answers to a multiple-choice question from a set of agents. Agents cannot communicate with each other during the elicitation process, and they do not know each other's identity. The requester then rewards agents based on their reported answers using our proposed scoring method. We assume that agents' rewards are somehow coupled with

relevant incentives, be they social-psychological, such as praise or visibility, or material rewards through prizes or money. Then, we naturally assume that agents seek to maximize their expected rewards. We show that, in the presence of social projection, agents strictly maximize their expected rewards by honestly disclosing their answers.

We also show how our proposed scoring method reduces to the *output-agreement method* whenever the underlying proper scoring rule is bounded and symmetric. The output-agreement method rewards agents based on whether their reported answers are the same. The output-agreement method has been successfully used in online marketplaces and in games with a purpose such as the ESP game [87]. We also show how our scoring method together with the ranked probability scoring rule [29] induce honest reporting by penalizing disagreements proportionally to the absolute difference between agents' answers.

From an empirical perspective, we investigate the efficiency of the output-agreement method and, consequently, of our scoring method in a content-analysis experiment. When objective verification is not possible, as often in content analysis, economic measures may be used to encourage agents to honestly disclose their answers. We discuss how our scoring method does so by making pairwise comparisons between reported answers and rewarding agreements. Our experimental results show that encouraging honest reporting through the output-agreement method creates more accurate answers than when agents have no direct incentives for expressing their honest answers.

3.1 Related Work

Two prominent methods to induce honest reporting without relying upon the existence of a ground-truth answer have been recently proposed: the *Bayesian truth serum* (BTS) *method* [65] and the *peer-prediction method* [54].

Similar to the setting we discuss in this chapter, the BTS method works on a single multiple-choice question with a finite number of answers. Each agent endorses the answer mostly likely to be correct and predicts the empirical distribution of the endorsed answers. Agents are evaluated by the accuracy of their predictions as well as based on how surprisingly common their answers are, *i.e.*, based on whether the reported answer is more common than collectively predicted.

The score received by an agent from the BTS method has two major components. The first one, called the information score, evaluates the agent's answer according to the log-ratio of its actual-to-predicted endorsement frequencies. The second component, called the prediction score, is a penalty proportional to the relative entropy between the empirical

distribution of answers and the agent's prediction of that distribution. Under the BTS scoring method, collective honest reporting is a Bayes-Nash equilibrium, *i.e.*, the best action that an agent can take (in expectation) given that all his peers are reporting honestly is also to report honestly.

The BTS method has been used to promote honest reporting in many different domains, *e.g.*, when sharing rewards amongst a set of agents [16] and in policy analysis [93]. However, there are two major drawbacks with the BTS method. First, it requires the population of agents to be large. Second, besides reporting their answers, agents must also make predictions about how their peers will report their answers. While the artificial intelligence community has recently addressed the former issue [67, 96], the latter issue is still an intrinsic requirement for using the BTS method.

The peer-prediction method [54], on the other hand, does not share the drawbacks of the BTS method. In the original setting of the peer-prediction method, a number of agents experience a product and rate its quality. A mechanism then collects the ratings and makes payments based on those ratings. The peer-prediction method makes use of the stochastic correlation between the signals observed by the agents from the product to achieve a Bayes-Nash equilibrium where every agent reports honestly.

Our work differs from the aforementioned works in that we explicitly explore the concept of *social projection* to induce honest reporting. Social projection is a cognitive bias where an agent believes that his private information is the most common one amongst his peers, although that might not be the case in reality. In other words, agents project themselves onto others. As the literature on psychological projection has shown, social projection serves as an egocentric heuristic for inductive reasoning [68]. There is evidence that social projection exists intra-groups and, to a less degree, inter-groups [68]. Some formal models have been proposed to model social projection, *e.g.*, Brenner and Bilgin [11] proposed a social projection model based on support theory, whereas Busemeyer and Pothos [13] discussed a quantum model of social projection.

We propose a flexible social projection model that allows one to model the strength of the projection within the Bayesian learning framework. This modeling choice is desirable because it has been shown that the strength of the projection is context dependent, *e.g.*, intra-group social projection seems to be stronger in groups that are artificially created in the laboratory than in groups that exist in the social world [68].

Given our social projection model, we show that it is possible to induce honest reporting by simply comparing reported answers and rewarding agreements. The rationale behind our approach is that when an agent believes that his answer is the most popular one, then honest reporting maximizes the chance of a random agreement. It is interesting to note

that both the BTS method and the peer-prediction method only induce honest reporting in a Bayes-Nash Equilibrium, whereas this is not the case for our method. The main reason for this result is that our model is simpler in that it only relies on an agent's belief about the population knowledge, instead of an agent's belief about others' beliefs about the population knowledge.

Rewarding agents based on pairwise comparisons has been empirically proven to be an effective payment structure in some domains. For example, Shaw *et al.* [75] measured the effectiveness of a collection of social and financial incentive schemes for motivating agents to conduct a qualitative content-analysis task. The authors found that treatment conditions that provided financial incentives and asked agents to prospectively think about the responses of their peers produced more accurate responses. Huang and Fu [39] showed that informing the agents that their rewards will be based on how similar their responses are to other agents' responses produces more accurate responses than telling the agents that their rewards will be based on how similar their responses are to gold-standard responses. The aforementioned studies were both performed on Amazon's Mechanical Turk.

Our work adds to the existing body of literature by providing a theoretical justification, and further empirical evidence, on why scoring methods based on pairwise comparisons work well in environments such as Amazon's Mechanical Turk. Moreover, our work provides a novel characterization of the output-agreement method different than, for example, a recent game-theoretic characterization by Waggoner and Chen [88]. Waggoner and Chen argued that the output-agreement method does not elicit honest answers. Instead, it elicits the correct answer according to the common knowledge among agents. We obtain a different result because, in contrast to Waggoner and Chen's work, we make assumptions on the nature of agents' information structure to model social projection.

3.2 The Model

We consider a multiple-choice question with a total of $n \geq 2$ exhaustive and mutually exclusive answers $A_1, \ldots, A_n$. This modeling choice is rather flexible because one can phrase many different tasks in terms of a multiple-choice question, *e.g.*, ratings, content/sentiment analysis, classification, categorization, *etc.* We assume that the population's knowledge is represented by an *unknown* categorical distribution Ω with parameter $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$, where $0 \leq \omega_k \leq 1$ and $\sum_{k=1}^{n} \omega_k = 1$. A possible interpretation of ω_k is that it is the probability that an agent selected at random from the population of agents has A_k as the answer to the multiple-choice question.

Each agent possesses a privately observed draw (signal) from Ω. We refer to observed signals as *honest answers*. We denote the honest answer of an agent i by $t_i \sim \Omega$, where $t_i \in \{A_1, \ldots, A_n\}$. Honest answers are independent, *i.e.*, $P(t_i|t_j) = P(t_i)$. Agents are not necessarily honest. For example, agents might report an answer they believe is socially desirable, as opposed to their true answers. Hence, we distinguish between each agent i's honest answer, t_i, and his reported answer $r_i \in \{A_1, \ldots, A_n\}$. We say that agent i is reporting honestly when his reported answer is equal to his honest answer, *i.e.*, $r_i = t_i$.

An entity, referred to as the *center*, is responsible for eliciting the answers and for rewarding the agents. Let s_i be agent i's *reward* after he reports r_i. We discuss how to compute s_i in the next section. Rewards are somehow coupled with relevant incentives, be they social-psychological, such as praise or visibility, or material rewards through prizes or money. We make four major assumptions in our model:

1. *Autonomy*: agents cannot influence other agents' answers, *i.e.*, they do not know each other's identity and they are not allowed to communicate to each other during the elicitation process.

2. *Risk Neutrality*: agents behave so as to maximize their expected rewards.

3. *Symmetric Dirichlet Priors*: each agent i has a prior distribution over $\boldsymbol{\omega}$. We assume that this prior is a symmetric Dirichlet distribution with hyperparameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, *i.e.*, $P(\boldsymbol{\omega}|\,\boldsymbol{\alpha})$.

4. *Social Projection*: after observing t_i, every agent i updates his belief by applying Bayes' rule to his prior, *i.e.*, $P(\boldsymbol{\omega}|\boldsymbol{\alpha}, t_i)$.

The first assumption means that agents work individually on the multiple-choice question. It describes how, for example, crowd workers traditionally solve tasks on the crowdsourcing platform Amazon's Mechanical Turk. The second assumption means that agents are self-interested and no external incentives exist for each agent.

The third assumption means that all the elements of the hyperparameter $\boldsymbol{\alpha}$ have the same value, *i.e.*, $\alpha_1 = \cdots = \alpha_n > 0$. In practice, this assumption implies that before an agent sees the underlying multiple-choice question and knowing only the number of possible answers n, the agent's response to the query "*what is the probability that one of your peers will report the answer A_k?*" is $\frac{1}{n}$, for all $k \in \{1, \ldots, n\}$. That is, all the answers are equally likely *a priori*. Formally:

$$P\left(r_j = A_k|\boldsymbol{\alpha}\right) = \mathbb{E}[\omega_k|\boldsymbol{\alpha}] = \frac{\alpha_k}{\sum_{x=1}^{n}\alpha_x} = \frac{1}{n}$$

for $j \neq i$. Given that all the answers are equally likely *a priori*, we henceforth refer to the prior as *uninformative*. The fourth assumption means that the posterior distributions are consistent with Bayesian updating. Moreover, this assumption implies that *after* solving the multiple-choice question, an agent's response to the query "*what is the probability that one of your peers will report the answer* A_k*?*" is:

$$P\left(r_j = A_k|\boldsymbol{\alpha}, t_i\right) = \mathbb{E}[\omega_k|\boldsymbol{\alpha}, t_i] = \begin{cases} \frac{\alpha_k+1}{1+\sum_{x=1}^{n}\alpha_x} & \text{if } t_i = A_k, \\ \frac{\alpha_k}{1+\sum_{x=1}^{n}\alpha_x} & \text{otherwise.} \end{cases}$$

We can write the above posterior as follows:

$$P\left(r_j = A_k|\boldsymbol{\alpha}, t_i\right) = \mathbb{E}[\omega_k|\boldsymbol{\alpha}, t_i] = \begin{cases} \frac{1}{n} + y & \text{if } t_i = A_k, \\ \frac{1}{n} - \frac{y}{n-1} & \text{otherwise.} \end{cases} \tag{3.1}$$

for $0 < y = \frac{n-1}{n+\alpha_k \times n^2} < 1$. Appendix A provides details about Dirichlet distributions. The above equation models *social projection*, which is the tendency to expect similarities between oneself and others [68]. In other words, an agent believes that his honest answer is the most popular answer amongst his peers. With this perspective, the value of y and, consequently, the elements $\alpha_1, \ldots, \alpha_n$ of the hyperparameter $\boldsymbol{\alpha}$ determine the strength of the social projection, where a small (respectively, high) value for y implies a weak (respectively, strong) social projection.

3.2.1 Social Projection

An important assumption in our model is that social projection holds true. Social projection, as defined in our work, is an extreme version of the psychological concept known as the *false-consensus effect* [69]. False consensus refers to an egocentric bias that occurs when people estimate consensus for their own behaviors or beliefs. Specifically, the false-consensus hypothesis asserts that human beings who engage in a certain behavior, or hold a certain belief, estimate that behavior/belief to be more common than the estimations from people who engage in alternative behaviors, or hold different beliefs.

The false-consensus effect is not restricted to cases where people believe that their values are shared by the majority. In other words, the term false consensus has also been used

when there is no consensus, *i.e.*, when agents do not necessarily believe that the majority of people share their views, but their estimates of the number of people who share their views tend to exceed the actual number. Our concept of social projection, on the other hand, asserts that an agent believes that his true response is the most popular response amongst his peers, *i.e.*, there is actually a consensus around the agent's true response, which might be a false consensus.

A pioneer work on false consensus was performed by Katz and Allport [44], who noticed that the more students admitted that they had cheated on an exam, the more they expected that other students cheated too. Ross *et al.* [69], the authors who coined the term false consensus, hypothesized and then demonstrated that people tend to overestimate the popularity of their own beliefs and preferences. In particular, the results of Studies 1, 3, and 4 in the paper by Ross *et al.* [69] showed that social projection, as defined in this paper, holds true when considering the agents' aggregated answers and estimates, whereas the combined results of Study 2 showed that social projection was present in 5 out of 14 multiple-choice questions.

Since the work by Ross *et al.* [69], dozens of studies have documented a systematic relationship between one's perceptions of his own characteristics and his estimates of the percentage of people in the population who share those characteristics in a variety of settings, ranging from questionnaire studies presenting situations, choices, and judgments that are hypothetical to actual conflict situations demanding personally relevant behavioral choices and social judgments [55, 51, 68]. Moreover, it has been shown that the strength of the false-consensus effect is highly dependent on the underlying agents and task. For example, there is some evidence that elderly people display higher degree of false consensus than adolescents [97]. Furthermore, false consensus is stronger when people make judgments about ingroups than when they make judgments about outgroups [68].

The emergency of the false-consensus effect and, consequently, our concept of social projection, has been explained under four different perspectives [51]: 1) selective exposure and cognitive availability; 2) salience and focus of attention; 3) motivational processes; and 4) logical information processing.

The perspective of selective exposure and availability suggests that instances of similarity between oneself and others are more readily available from memory than instances of dissimilarity, thereby increasing estimates of consensus for one's preferred opinion. The reason why these instances of similarity are more readily available is due to selective exposure, *i.e.*, due to the tendency of human beings to associate with those who are similar rather than dissimilar to themselves. This selective exposure to similar others provides an agent with a biased and restricted sample of information about the population's true

diversity of opinion.

The perspective about salience and focus of attention suggests that false consensus/social projection arises when an agent focuses attention exclusively on a single position. Consequently, perceived consensus may be augmented because that position is the only one in the agent's immediate consciousness.

The motivational perspective asserts that false consensus/social projection has a dissonance-reduction function, where agents unconsciously project their own beliefs onto others in order to obtain confirmation for their own attitudes and beliefs, *i.e.*, to bolster social support, to validate the correctness or appropriateness of a belief, to maintain self-esteem, *etc.*

Finally, the logical information processing perspective views active reasoning and rational processes as underlying one's estimates about the similarity between oneself and others. Within this perspective, Dawes [26] reexamined the data obtained by Ross *et al.* [69] and argued, from a Bayesian perspective, that subjects were correct in considering their own behavioral choices common in the population. In particular, Dawes [26] showed that the typical strength of the false-consensus effect was similar to the statistically normative change from prior to posterior probabilities in generic induction. Moreover, Dawes [26] discussed that even a sample of size 1, representing an agent's private information as in our model, should have substantial effect on the agent's percentage estimates. Therefore, it is conceivable that subjects intuitively understand the logic of statistical induction and perform accordingly.

To summarize, our concept of social projection is strongly grounded in psychology. In particular, social projection is an extreme version of the psychological concept known as the false-consensus effect. As the seminal work by Ross *et al.* [69] suggests, social projection happens often and in many different settings.

3.3 The Scoring Method

If the center knew *a priori* agents' honest answers, he could then compare the honest answers to the reported answers and reward agreement. However, due to the subjective nature of our setting, we face a situation where this objective truth is unknowable. Our solution to this issue is to induce honest reporting by providing *rewards* based on pairwise comparisons between reported answers.

The first step towards computing an agent i's reward is to estimate his posterior distribution in (3.1) based on his reported answer r_i. Let $\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i] = (\mathbb{E}[\omega_1|\boldsymbol{\alpha}, r_i], \ldots, \mathbb{E}[\omega_n|\boldsymbol{\alpha}, r_i])$

be such an estimation, where:

$$\mathbb{E}[\omega_k|\boldsymbol{\alpha}, r_i] = \begin{cases} \frac{1}{n} + y & \text{if } r_i = A_k, \\ \frac{1}{n} - \frac{y}{n-1} & \text{otherwise.} \end{cases} \tag{3.2}$$

Clearly, $\mathbb{E}\left[\omega_k|\boldsymbol{\alpha}, r_i\right] = \mathbb{E}\left[\omega_k|\boldsymbol{\alpha}, t_i\right]$ if and only if agent i is reporting honestly, *i.e.*, when he reports $r_i = t_i$. The score of agent i is then determined as follows:

$$s_i = \gamma R\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right) + \lambda \tag{3.3}$$

where $j \neq i$ is an arbitrary agent, γ and λ are constants, for $\gamma > 0$ and $\lambda \in \mathbb{R}$, and R is a strictly proper scoring rule. Scoring rules require an observable outcome, or a ground truth, in order to score an assessment. Intuitively, the scoring method in (3.3) considers the answer reported by an agent other than agent i as the observed outcome, and then the method scores agent i's *estimated posterior distribution* in (3.2) as an assessment of the observed outcome. In the following proposition, we show that our scoring method induces honest reporting.

Proposition 2. *Each agent i strictly maximizes his expected reward if and only if $r_i = t_i$.*

Proof. For ease of exposition, let $\boldsymbol{\Theta}_i = \mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, t_i\right]$ and $\boldsymbol{\Phi}_i = \mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right]$. By the autonomy assumption, agent j cannot affect agent i's reported answer.

(If part) Since R is a strictly proper scoring rule, from Proposition 1 we have that:

$$\arg\max_{\boldsymbol{\Phi}_i} \mathbb{E}_{\boldsymbol{\Theta}_i}\left[\gamma R\left(\boldsymbol{\Phi}_i, r_j\right) + \lambda\right] = \boldsymbol{\Theta}_i$$

If $r_i = t_i$, then by construction $\boldsymbol{\Phi}_i = \boldsymbol{\Theta}_i$, *i.e.*, the estimated posterior distribution in (3.2) is equal to the true posterior distribution in (3.1). Consequently, honest reporting maximizes agents' expected rewards.

(Only-if part). Using a similar argument, given that R is a strictly proper scoring rule, from Proposition 1 we have that:

$$\arg\max_{\boldsymbol{\Phi}_i} \mathbb{E}_{\boldsymbol{\Theta}_i}\left[\gamma R\left(\boldsymbol{\Phi}_i, r_j\right) + \lambda\right] = \boldsymbol{\Theta}_i$$

By construction, $\boldsymbol{\Phi}_i = \boldsymbol{\Theta}_i$ if and only if $r_i = t_i$. Thus, agents maximize their expected rewards only when they are honest. □

Another way to interpret Proposition 2 is to imagine that each agent is betting on the most appropriate answer to the underlying multiple-choice question. Since the most relevant information available to him is the observed signal, then the strategy that maximizes his expected reward is to bet on that signal, *i.e.*, to bet on his honest answer. When this happens, the true posterior distribution in (3.1) is equal to the estimated posterior distribution in (3.2) and, consequently, the expected score resulting from a strictly proper scoring rule is strictly maximized when the expectation is taken with respect to the true posterior distribution.

We note that because honest answers are independent, Proposition 2 is still valid when the scoring method in (3.3) involves more than one agent, *i.e.*:

$$s_i = \sum_{j \neq i} \left(\gamma R \left(\mathbb{E} \left[\boldsymbol{\omega} | \boldsymbol{\alpha}, r_i \right], r_j \right) + \lambda \right) \tag{3.4}$$

Using only a few agents in the above summation might result in "unlucky" pairwise comparisons, *i.e.*, a single agent with a very distinct reported answer might cause another agent to have a low reward. Hence, we believe there are practical benefits in using many agents in order to obtain fairer rewards. It is interesting to compare our scoring method with a related method known as the Bayesian Truth Serum (BTS) method. Using our model, the BTS method can be defined as:

$$\sum_{k=1}^{n} h(r_i, A_k) \ln \frac{\bar{A}_k}{\bar{\phi}_k} + \sum_{k=1}^{n} \bar{A}_k \ln \frac{\phi_{i,k}}{\bar{A}_k},$$

where $h(\cdot)$ is the identity function, $\bar{A}_k$ is the average frequency of answer A_k, $\phi_{i,k}$ is agent i's reported probability regarding the occurrence of answer A_k, and $\bar{\phi}_k$ is the geometric average of the predicted frequencies of answer A_k. As mentioned before, collective honest reporting is a Bayes-Nash equilibrium under the BTS method. As we will show in the following subsections, our method can be reduced to a much simpler and intuitive scoring function than the BTS method.

Different interpretations of the scoring method in (3.3) may arise depending on the underlying strictly proper scoring rule. In the following subsections, we discuss two natural interpretations for the proposed scoring method: 1) when R is a symmetric and bounded strictly proper scoring rule; and 2) when R is a strictly proper scoring rule sensitive to distance called *ranked probability scoring rule*.

3.3.1 Rewarding Agreements

When agents' prior distributions are uninformative, the elements of agents' true and estimated posterior distributions can take on only two possible values (see equations (3.1) and (3.2)). Consequently, if R is a symmetric scoring rule, then the term $R\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right)$ in (3.3) can take on only two possible values because a permutation of similar probability values in $\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right]$ does not change the score from a symmetric scoring rule. When R is also strictly proper, it means that $R\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right) = \delta_{max}$, when $r_i = r_j$, and $R\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right) = \delta_{min}$, when $r_i \neq r_j$, where $\delta_{max} > \delta_{min}$. Thus, we can write (3.3) as follows:

$$\gamma R\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right) + \lambda = \begin{cases} \gamma\delta_{max} + \lambda & \text{if } r_i = r_j, \\ \gamma\delta_{min} + \lambda & \text{otherwise.} \end{cases}$$

When R is also bounded, we can then set $\gamma = \frac{1}{\delta_{max} - \delta_{min}}$ and $\lambda = \frac{-\delta_{min}}{\delta_{max} - \delta_{min}}$, and the above values become, respectively, 1 and 0. Hence, the resulting scores do not depend on parameters of the model. Moreover, we obtain an intuitive interpretation of the scoring method in (3.3), namely whenever two reported answers are equal to each other, the underlying agents are rewarded by one payoff unit. This is precisely what constitutes the *output-agreement method* [87, 88, 14]. In other words, the output-agreement method can be seen as an instance of our scoring method and, consequently, it induces honest reporting in our setting. This result may explain why the output-agreement method has been successfully used in practice.

3.3.2 Taking Distance Into Account

Rewarding agreements, as defined in the previous subsection, might work well for small values of n, the total number of answers, but it can be too restrictive and, to some degree, unfair when n is high. For example, consider the case when the answers are numerical values, where $A_1 = 1, A_2 = 2, \ldots, A_n = n$. Moreover, let $n = 5$, and the answer used as the observed outcome equal to $r_j = A_5 = 5$. Then, a reported answer $A_4 = 4$ seems to be more accurate than a reported answer equal to $A_1 = 1$. One effective way to deal with this issue is by using a strictly proper scoring rule in (3.3) that is *sensitive to distance*.

Using the notation of Section 2.1, recall that $\mathbf{q} = (q_1, \ldots, q_n)$ is some probability vector. We assume that the outcomes (answers) are ordered numerical values, *i.e.*, $A_1 = 1 < A_2 = 2 < \cdots < A_n = n$. We denote the cumulative probabilities by capital letter: $Q_k = \sum_{j \leq k} q_j$. We first define the notion of distance between two probability vectors as proposed by Staël

von Holstein [79]. We say that a probability vector $\mathbf{q}'$ is more distant from the jth outcome than a probability vector $\mathbf{q} \neq \mathbf{q}'$ if:

$$Q'_k \geq Q_k, \text{ for } k = 1, \ldots, j-1$$
$$Q'_k \leq Q_k, \text{ for } k = j, \ldots, n$$

Intuitively, the above definition means that $\mathbf{q}$ can be obtained from $\mathbf{q}'$ by moving probability mass towards the jth outcome from other outcomes [79]. A scoring rule R is said to be *sensitive to distance* if $R(\mathbf{q}, A_j) > R(\mathbf{q}', A_j)$ whenever $\mathbf{q}'$ is more distant from $\mathbf{q}$ for all outcomes A_j. Epstein [29] introduced the *ranked probability score* (RPS), a strictly proper scoring rule that is sensitive to distance. Using the formulation of Epstein's result proposed by Murphy [56], we have for a probability vector $\mathbf{q}$ and an observed outcome A_j:

$$RPS(\mathbf{q}, A_j) = -\sum_{k=1}^{j-1} Q_k^2 - \sum_{k=j}^{n} (1 - Q_k)^2 \tag{3.5}$$

When using RPS as the strictly proper scoring rule in (3.3), the center rewards agents based on how close their reported answers are to the answers taken as observed outcomes. We show below that this closeness concept has a very natural interpretation in our model. Assume that the elements of the hyperparameter $\boldsymbol{\alpha}$ are defined as follows: $\alpha_1 = \cdots = \alpha_n = \epsilon$, for an arbitrarily small constant $\epsilon > 0$. In this case, each element of the posterior distribution in (3.1) is:

$$\mathbb{E}\left[\omega_k | \boldsymbol{\alpha}, t_i\right] = \begin{cases} \frac{\epsilon+1}{1+\sum_{x=1}^{n} \epsilon} & \text{if } t_i = A_k, \\ \frac{\epsilon}{1+\sum_{x=1}^{n} \epsilon} & \text{otherwise.} \end{cases}$$

For a sufficiently small ϵ, the above value is approximately:

$$\mathbb{E}\left[\omega_k | \boldsymbol{\alpha}, t_i\right] = \begin{cases} 1 & \text{if } t_i = A_k, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, each element of the estimated posterior distribution in (3.2) is approximately:

$$\mathbb{E}\left[\omega_k | \boldsymbol{\alpha}, r_i\right] = \begin{cases} 1 & \text{if } r_i = A_k, \\ 0 & \text{otherwise.} \end{cases}$$

Under the above circumstances, when R in (3.3) is the ranked probability scoring rule, the term $R\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right)$ becomes:

$$\begin{aligned} RPS\left(\mathbb{E}\left[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i\right], r_j\right) &= -\sum_{k=1}^{r_j-1}\left(\sum_{x\leq k}\mathbb{E}\left[\omega_x|\boldsymbol{\alpha}, r_i\right]\right)^2 - \sum_{k=r_j}^{n}\left(1-\sum_{x\leq k}\mathbb{E}\left[\omega_x|\boldsymbol{\alpha}, r_i\right]\right)^2 \\ &= -\left|r_i - r_j\right| \end{aligned}$$

In words, scores from R are equal to the negative of the absolute difference between two reported answers. Consequently, the scoring method in (3.3) has a very natural interpretation: agents are penalized in proportion to the absolute difference between their reported answers. Since the scoring range of RPS is $[-n+1, 0]$, the resulting scores are always non-negative when $\gamma = 1$ and $\lambda = n - 1$.

3.3.3 Numerical Example

Consider a population of four agents and a multiple-choice question with five numerical answers: $A_1 = 1, A_2 = 2, A_3 = 3, A_4 = 4$, and $A_5 = 5$. Suppose that agents have uninformative Dirichlet priors with $\boldsymbol{\alpha} = (\epsilon, \epsilon, \epsilon, \epsilon, \epsilon)$, where $\epsilon > 0$ is an arbitrarily small constant. Assume that agents 1, 2, 3, and 4 report, respectively, $r_1 = 1$, $r_2 = 1$, $r_3 = 2$, and $r_4 = 5$. From (3.2), the resulting estimated posterior distributions are, approximately, $\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_1 = 1] = (1, 0, 0, 0, 0)$, $\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_2 = 1] = (1, 0, 0, 0, 0)$, $\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_3 = 2] = (0, 1, 0, 0, 0)$, and $\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_4 = 5] = (0, 0, 0, 0, 1)$. In this example, agents' information structures are entirely defined by their observed signals. In what follows, we illustrate the scores the proposed scoring method returns when using a symmetric and bounded strictly proper scoring rule and when using RPS. These scores show how distinct reported answers might have a considerable impact on agents' scores when the center does not take distance into account.

Rewarding Agreements

Assume that R in (3.3) is the quadratic scoring rule in (2.2), which in turn is symmetric, bounded, and strictly proper. Consequently, as discussed in Section 3.3.1, the term $\gamma R\left(\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, r_i], r_j\right) + \lambda$ in (3.4) can take on only two values:

$$\begin{array}{ll} \gamma\left(2\times 1-1^2-\sum_{x=1}^{n-1}0^2\right)+\lambda & \text{if } r_i=r_j, \\ \gamma\left(2\times 0-1^2-\sum_{x=1}^{n-1}0^2\right)+\lambda & \text{otherwise.} \end{array}$$

If we set $\gamma=\frac{1}{\delta_{max}-\delta_{min}}=0.5$ and $\lambda=\frac{-\delta_{min}}{\delta_{max}-\delta_{min}}=0.5$, for $\delta_{max}=1$ and $\delta_{min}=-1$, the above values are equal to, respectively, 1 and 0. Using the scoring method in (3.4), we then obtain the following scores: $s_1=s_2=1$ and $s_3=s_4=0$. That is, the rewards agents 1 and 2 receive are similar due to the fact that $r_1=r_2$. Agents 3 and 4's rewards are equal to 0 because there is no match between their reported answers and others' reported answers.

Taking Distance into Account

Now, assume that R in (3.4) is the ranked probability score shown in (3.5). In order to ensure non-negative scores, let $\gamma=1$ and $\lambda=n-1=4$. Using the scoring method in (3.4), we obtain the following rewards: $s_1=s_2=(\lambda-|1-1|)+(\lambda-|1-2|)+(\lambda-|1-5|)=7$, $s_3=2\times(\lambda-|2-1|)+(\lambda-|2-5|)=7$, and $s_4=2\times(\lambda-|5-1|)+(\lambda-|5-2|)=1$. The reward of agent 4 is the lowest because his reported answer is the most different answer, *i.e.*, it has the largest distance between it and all of the other answers.

3.4 Aggregating Reported Answers

After agents report their answers and receive their rewards, there is still the question of how the center will aggregate the reported answers. We explain in this section why the center can rely on the frequency of the reported answers when he induces honest reporting.

For illustration's sake, assume that the center and the agents have common uninformative Dirichlet priors, where the elements of the hyperparameter $\boldsymbol{\alpha}$ are defined as follows: $\alpha_k=\epsilon$, for all $k\in\{1,\ldots,n\}$, and $\epsilon>0$ is an arbitrarily small constant. Moreover, assume that the center updates his belief using Bayes' rule after observing the reported answers, *i.e.*, $P(\boldsymbol{\omega}|\boldsymbol{\alpha},r_1,\ldots,r_n)$. Let h_k be the number of times that the answer $A_k\in\{A_1,\ldots,A_n\}$ is reported. Hence, the center's final posterior distribution is approximately:

$$\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha},r_1,r_2,\ldots,r_n]=\left(\frac{h_1}{\sum_{x=1}^{n}h_x},\frac{h_2}{\sum_{x=1}^{n}h_x},\ldots,\frac{h_n}{\sum_{x=1}^{n}h_x}\right)$$

In other words, the center's final posterior distribution is equal to the frequency of the reported answers. Due to the law of large numbers, the frequency of the reported answers converges to the distribution that represents the population knowledge as the number of honestly reported answers increases. This result, together with the result from Section 3.3, provide a theoretical justification for why the output-agreement method and, more generally, our proposed scoring method is expected to perform well in the studied setting. This section's result also means a word of caution on indiscriminately using the frequency of agents' reports as the distribution of agents' observed signals. Such a perspective is unlikely to hold true when agents are not reporting their private information honestly.

3.5 Experiments

In this section, we describe a content-analysis experiment designed to test the effectiveness of our scoring method. In the following subsections, we discuss the experimental design and the experimental results.

3.5.1 Experimental Design

We designed a content-analysis task on Amazon's Mechanical Turk (AMT) that required workers (agents) to review 3 short texts under three different criteria: *Grammar*, *Clarity*, and *Relevance*. The first two texts were extracts from published poems, but with some original words intentionally replaced by misspelled words. The third text contained random words presented in a semi-structured way. Appendix B contains all the details regarding the texts. For each text, three questions were presented to the agents, each one having three possible answers ordered in decreasing negativity order:

- Grammar: does the text contain misspellings, syntax errors, *etc.*?
  - A lot of grammar mistakes
  - A few grammar mistakes
  - No grammar mistakes
- Clarity: does the text, as a whole, make any sense?
  - The text does not make sense

  - The text makes some sense
  - The text makes perfect sense

- Relevance: could the text be part of a poem related to love?
  - The text cannot be part of a love poem
  - The text might be part of a love poem
  - The text is definitely part of a love poem

We intentionally designed subjective answers so as to emphasize the subjective nature of content analysis. We translated each reported answer into a numeral value inside the set $\{0, 1, 2\}$. The most negative answer received the score 0, the middle answer received the score 1, and the most positive answer received the score 2. Thus, each agent reported a vector of 9 numerical values (3 texts times 3 criteria).

We recruited 150 agents on AMT, all of them residing in the United States of America and older than 18 years old. They were required to accomplish the content-analysis task in at most 20 minutes. We split the agents into 3 groups of equal size. Agents were not informed about the group sizes. After accomplishing the task, every agent in every group received a payment of \$0.20. A study done by Ipeirotis [40] showed that more than 90% of the tasks on AMT have a baseline payment less than \$0.10, and 70% of the tasks have a baseline payment less than \$0.05. Thus, our baseline payment was much higher than the majority of payments from other tasks posted on the Amazon's Mechanical Turk marketplace.

We randomly assigned each agent into one of the three groups. Agents in two of the groups, the treatment groups, could earn an additional *bonus* of up to \$0.10. We informed agents in the first treatment group, referred to as the *Bonus Group* (BG), that their bonuses would be proportional to the number of answers similar to their reported answers. Agents in the second treatment group, the *Bonus and Information Group* (BIG), received similar information, but they also received a short summary of some theoretical results presented in this paper:

> "A group of researchers from the University of Waterloo (Canada) formally showed that the best strategy to maximize your expected bonus in this setting is by being honest, *i.e.*, by considering each question thoroughly and deciding the best answer according to your personal opinion".

The main reason behind using the treatment group BIG was to investigate the effects of framing on the accuracy of the reported answers. Members of the third group, the *Control Group* (CG), neither received extra explanations nor bonuses. We used their reported answers as the control condition. We computed the bonuses by rewarding agreements as described in Section 3.3.1. Recall that each agent reported 9 answers (3 texts times 3 criteria). For each answer reported by an agent i, we calculated the total number of agreements (#agreements) between agent i's reported answers and the answers reported by members of agent i's group. Consequently, for each reported answer, there could be at most 49 similar reported answers because each group had 50 members. We then used the formula $\frac{10}{9} \times \frac{\#\text{agreements}}{49}$ to calculate the reward for an individual answer. Such a reward can be seen as a positive affine transformation of the scoring method in (3.4) and, consequently, it induces honest reporting due to Proposition 1. Given that each agent reported 9 answers, if the answers reported by all members of a group were the same, then all group members would received the maximum bonus of 10 cents.

3.5.2 Gold-Standard Answers

We were able to derive *gold-standard answers* for each text/criterion because we knew the source and original content of each text *a priori*, *i.e.*, before we conducted the content-analysis experiment. In order to avoid confirmation bias[1], we asked five professors and tutors from the English and Literature Department at the University of Waterloo to complete the content-analysis task. We set the *gold-standard answer* for each criterion in a text as the median of the answers reported by the professors and tutors. Coincidentally, each median was also the mode of the underlying answers. Appendix B contains all the answers reported by the professors and tutors as well as the respective gold-standard answers.

3.5.3 Hypotheses

Our first research question was whether or not providing scores through pairwise comparisons makes the reported answers more accurate, *i.e.*, closer to the gold-standard answers. Our hypothesis was:

Hypothesis 1. *The average error of group BIG is less than the average error of group BG, which in turn is less than the average error of group CG.*

[1]The tendency to interpret information in a way that confirms one's preconceptions [64].

In other words, the resulting answers would be on average more accurate when agents' bonuses were based on our scoring method. Moreover, the extra explanation regarding the theory behind the scoring method would provide more credibility to it, thus making the reported answers more accurate. That is, framing has a positive influence on the accuracy of the reported answers. Regarding the resulting bonuses, since honest reporting maximizes agents' expected scores in our model (Proposition 2), our second hypothesis was:

Hypothesis 2. *The average bonus received by members of group BIG is greater than the average bonus received by members of group BG, which in turn is greater than the average bonus received by members of group CG.*

In order to test whether Hypothesis 2 was true, we used the bonus the members of group CG would have received had they received any bonus. It is noteworthy that we measured Hypothesis 1 by comparing how close the reported answers were to the gold-standard answers, whereas we measured Hypothesis 2 by making pairwise comparisons between reported answers: the higher the number of agreements, the greater the resulting bonus.

Another metric we used to compare groups' performance was the *task completion time*. The amount of time agents spent on the content-analysis task can be seen as a proxy for the effort they exerted to complete the task. Regarding this metric, we expected agents who received scores from our scoring method to be more cautious when completing their tasks. Moreover, the extra explanation regarding the theory behind the scoring method would provide more credibility to it, thus making the members of group BIG work longer on the task. Hence, our third hypothesis was:

Hypothesis 3. *The average task completion time of group BIG is greater than the average task completion time of group BG, which in turn is greater than the average task completion time of group CG.*

3.5.4 Experimental Results

In what follows, we describe our experimental results and analyze our hypotheses.

Error on Individual Criteria

In our first analysis, we defined the error of each agent's reported answer as the absolute difference between his answer and the corresponding gold-standard answer. Thus, the

outcome measure was an integer with a value between zero and two, and the closer this value was to zero, the better the resulting accuracy. Table 3.1 shows the average error for each group.

Table 3.1: The average of the absolute difference between the reported answers and the corresponding gold-standard answers is shown below each group. For each criterion, the lowest average is highlighted in bold. The standard deviations are in parenthesis. The last three columns show one-tailed p-values resulting from rank-sum tests. Given the notation A-B, the null hypothesis is that the outcome measures resulting from groups A and B are equivalent, and the alternative hypothesis is that the outcome measure resulting from group A is less than the outcome measure resulting from group B.

					p-values		
		BG	BIG	CG	BIG-BG	BIG-CG	BG-CG
	Grammar	0.5000	**0.3200**	0.4400	0.035**	0.110	0.726
		(0.5051)	(0.4712)	(0.5014)			
Text 1	Clarity	0.8200	**0.6200**	0.8600	0.065*	0.052*	0.413
		(0.6606)	(0.6024)	(0.7287)			
	Relevance	0.2200	**0.2000**	0.3000	0.484	0.213	0.230
		(0.5067)	(0.4518)	(0.5803)			
	Grammar	0.4400	**0.3600**	0.3800	0.209	0.420	0.729
		(0.5014)	(0.4849)	(0.4903)			
Text 2	Clarity	0.5000	**0.3800**	0.5400	0.155	0.067*	0.325
		(0.6468)	(0.6024)	(0.6131)			
	Relevance	**0.4400**	0.6400	0.6600	0.977	0.419	0.014**
		(0.5014)	(0.4849)	(0.4785)			
	Grammar	**0.7600**	0.7800	1.0200	0.539	0.077*	0.061*
		(0.8466)	(0.8640)	(0.8449)			
Text 3	Clarity	0.1400	**0.0000**	0.1600	0.006**	0.002**	0.301
		(0.4046)	(0.0000)	(0.3703)			
	Relevance	0.1200	**0.1000**	0.2000	0.491	0.112	0.122
		(0.4352)	(0.3642)	(0.4949)			

* $p \leq 0.1$
** $p \leq 0.05$

Focusing first on the groups BG and BIG, the group BIG is the most accurate group on all criteria, except for the criterion Relevance in Text 2 and the criterion Grammar in Text 3. This result is statistically significant with p-value ≤ 0.1 in three out of the seven cases in which BIG is more accurate than BG. In two out of these three statistically significant cases, this result is also statistically significant with p-value ≤ 0.05. BG is more accurate than BIG in only two criteria. This result is only statistically significant for the criterion Relevance in Text 2 (p-value ≤ 0.05). Hence, the priming effect has a positive impact on the accuracy of the reported answers.

The group CG, the control condition that involved no incentives beyond the baseline compensation offered for completing the content-analysis task, never outperforms the group BIG. Moreover, the group CG outperforms the group BG in only two occasions. In two (respectively, four) occasions, CG is statistically significantly less accurate than BG (respectively, BIG) with p-value ≤ 0.1.

Giving these results, we conclude that Hypothesis 1 is true for individual criteria, *i.e.*, the resulting answers are on average more accurate when using rewards based on pairwise comparisons, and the extra explanation regarding the theory behind our scoring method seems to provide more credibility to it, thus reducing the error of the reported answers.

Aggregate Error

We also computed the aggregate error for each text as well as for the whole task. In the former case, the outcome measure was the sum of the absolute differences between each reported answer for a given text and the corresponding gold-standard answer. For example, given $(0, 1, 2)$ as the reported answers for Text 1, and $(1, 2, 2)$ as the corresponding gold-standard answers, the outcome measure for Text 1 would be $|0-1|+|1-2|+|2-2| = 2$. For the whole task, we summed the absolute differences across all criteria and texts. Table 3.2 shows the average aggregate error for each group.

For every single text as well as for the overall task, members of the group CG report less accurate answers than members of the group BG and the group BIG. For the group BG, this result is statistically significant for Text 3 and for the overall task (p-value ≤ 0.1). For the group BIG, this result is statistically significant for Text 1 (p-value ≤ 0.1), Text 3 (p-value ≤ 0.05), and for the whole task (p-value ≤ 0.05). Thus, the experimental results suggest that providing rewards produces a significant improvement in quality over the control condition. Moreover, providing an extra explanation about the theory behind the scoring method improves the final quality of the answers because, on average, the answers from group BIG are more accurate than the answers from group BG. This result is

Table 3.2: The average of the sum of the absolute differences between the reported answers and the corresponding gold-standard answers is shown below each group. For each text and for the whole task, the lowest average is highlighted in bold. The standard deviations are in parenthesis. The last three columns show one-tailed p-values resulting from rank-sum tests. Given the notation A-B, the null hypothesis is that the outcome measures resulting from groups A and B are equivalent, and the alternative hypothesis is that the outcome measure resulting from group A is less than the outcome measure resulting from group B.

				p-values		
	BG	BIG	CG	BIG-BG	BIG-CG	BG-CG
Text 1	1.5400	**1.1400**	1.6000	0.043**	0.085*	0.588
	(1.1287)	(1.0304)	(1.4142)			
Text 2	**1.3800**	**1.3800**	1.5800	0.547	0.163	0.148
	(1.0669)	(0.9666)	(0.9916)			
Text 3	1.0200	**0.8800**	1.3800	0.394	0.020**	0.052*
	(1.1865)	(0.9179)	(1.1933)			
Overall	3.9400	**3.4000**	4.5600	0.110	0.002**	0.064*
	(2.2352)	(1.6903)	(2.1301)			

* $p \leq 0.1$
** $p \leq 0.05$

statistically significant for Text 1 (p-value ≤ 0.05). Therefore, we conclude that Hypothesis 1 is also true on the aggregate level.

Bonus

The average bonus per group is shown in the first row of Table 3.3. From it, we conclude that Hypothesis 2 is true, *i.e.*, the average bonus received by members of the group BIG is greater than the average bonus received by members of the group BG, which in turn is greater than the average bonus hypothetically received by members of the group CG. All these results are statistically significant with p-value ≤ 0.05. In other words, providing scores and informing agents about the theory behind the scoring method do indeed increase the number of reported answers that are similar.

Interestingly, there is a strong negative correlation between bonuses and the aggregate error for the whole task shown in the seventh row of Table 3.2, even though the former is computed by making pairwise comparisons between reported answers, whereas the latter

Table 3.3: Average bonus and completion time per group. The highest average values are highlighted in bold. The standard deviations are in parenthesis. The last three columns show one-tailed p-values resulting from rank-sum tests. Given the notation A-B, the null hypothesis is that the outcome measures resulting from groups A and B are equivalent, and the alternative hypothesis is that the outcome measure resulting from group A is greater than the outcome measure resulting from group B.

				p-values		
	BG	BIG	CG	BIG-BG	BIG-CG	BG-CG
Bonus	0.053	**0.058**	0.050	< 0.0005**	<0.0005**	0.0025**
	(0.0086)	(0.0073)	(0.0078)			
Time	178.66	**215.90**	196.36	0.0232**	0.0257**	0.4208
	(87.4495)	(127.7471)	(149.0788)			

** $p \leq 0.05$

is computed by comparing reported answers with gold-standard answers. The Pearson correlation coefficients for BG, BIG, and CG are, respectively, -0.73, -0.79, and -0.72. This result implies that there exists a strong positive correlation between honest reporting and accuracy in our content-analysis experiment.

Completion Time

We show the average completion time per group in the third row of Table 3.3. We start by noting that Hypothesis 3 is not true. Surprisingly, the average time spent on the task by members of the group BG is statistically equivalent to the average time spent by members of the group CG since the null hypothesis cannot be rejected. The average completion time by members of the group BIG is the highest one amongst the three groups, and this result is statistically significant with p-value ≤ 0.05. A possible explanation for this result is that agents work on the content-analysis task more seriously by taking more time to complete it when they receive a brief explanation regarding some theoretical results of the proposed scoring method, whereas the same agents could be quickly guessing how their peers would report their answers when the extra explanation about the theoretical results is not provided.

It is noteworthy that even though the average values might suggest that spending more time on the task results in higher bonuses and lower aggregate errors, we do not find any significant correlation between these variables at an individual level.

3.6 Summary

We proposed in this chapter a scoring method built on strictly proper scoring rules that induces honest reporting by risk-neutral agents in multiple-choice questions without relying upon the assumption of a ground-truth answer. We discussed two natural interpretations of our scoring method. First, we showed that the proposed scoring method reduces to the output-agreement method when the underlying proper scoring rule is bounded and symmetric. Thus, we provided a novel characterization of the output-agreement method. Second, we showed how our scoring method in conjunction with the ranked probability scoring rule induce honest reporting by penalizing disagreements proportionally to the absolute difference between reported answers.

We then discussed how honest reporting is related to accuracy in our model: as the number of honestly reported answers increases, the frequency of the reported answers convergences to the distribution that represents the population knowledge. Thus, the combination of our results explains why the output-agreement method and, more generally, our scoring method are expected to be effective in tasks framed as multiple-choice questions.

We tested the effectiveness of our scoring method on a content-analysis experiment using Amazon's Mechanical Turk. We empirically showed that providing rewards through pairwise comparisons results in more accurate answers than when agents have no direct incentives for expressing their honest answers. Moreover, agents tended to agree more with each other when they received such rewards.

3.6.1 Directions for Future Research

Given the positive results obtained in our content-analysis experiment, one interesting open question is whether our scoring method would perform as well in other domains. It has already been reported that the output-agreement method, which is an instance of our scoring method, is very effective in games with a purpose and crowdsourcing [87]. One particularly exciting domain for investigation is the peer-review process as used in massive open online courses since a peer-review task can be modeled as a multiple-choice question.

An important assumption in our model is the existence of social projection, *i.e.*, that agents assume that others will behave as they themselves do. As the literature on psychological projection has shown, social projection serves as an egocentric heuristic for inductive reasoning [68]. Consequently, social projection might explain why the output-agreement method has been so effective in crowdsourcing settings. An exciting direction for future research to validate or invalidate this hypothesis consists in determining the extent to which

social projection exists in crowdsourcing settings, and whether or not an equation similar to equation (3.1) can effectively model social projection.

Another question worth contemplating is whether or not incentives other than from the received rewards play a role in inducing honest behavior. For example, one can conjecture that altruism may play an important role in our scoring method. In our content-analysis experiment, for example, the performance of the agents not only affect their own rewards, but also the rewards of their peers. In other words, if agents do not put enough effort into reporting high-quality answers, not only might they receive low rewards, but other answers evaluated based on those erroneous answers might also receive low rewards. Thus, an interesting future work is to investigate whether or not agents, in general, have an altruistic motive to put more effort into the underlying task in order to maximize the potential payoffs of their peers.

Chapter 4

On Proper Scoring Rules and Cumulative Prospect Theory

An agent's assessment of the likelihood of a future event in which he has no stake may be of interest to others, *e.g.*, a financial investor may be interested in the probability a market expert assigns to the increase of a certain stock price. In the medical domain, a patient might want to know the likelihood of success of a treatment before deciding whether to undergo that treatment.

As described in Section 2.1, *proper scoring rules* are traditional scoring methods to induce honest reporting of subjective probabilities [94]. The expected score from a proper scoring rule is maximized when an agent reports his true belief. Hence, the implicit assumption behind proper scoring rules is that agents are *risk neutral*, *i.e.*, that they behave so as to maximize their expected scores.

Our first contribution in this chapter is an experiment where we investigate whether or not risk neutrality holds true when eliciting beliefs under proper scoring rules. Our experimental results show that human beings are not risk neutral when reporting their beliefs even when the stakes are small. We then discuss how reporting a belief under a proper scoring rule is equivalent to making a choice under uncertainty. Hence, our logical next step is to adapt proper scoring rules to a more appropriate model of choice under uncertainty[1].

Modern theories of individual choices under uncertainty based on rank-dependent util-

[1]We use the term *uncertainty* as commonly used in decision analysis, meaning that the uncertainty regarding the occurrence of future evens is captured by subjective probability distributions.

ities[2] assert that the values that human beings derive from payoffs are represented by nonlinear *utility functions*, which in turn are weighted by *decision weights*, instead of subjective probabilities [66, 73]. Decision weights are differences between *weighting functions* applied to cumulative probabilities. Thus, according to rank-dependent models, the components that drive an agent's attitude towards uncertainty are both a utility function and weighting functions.

Cumulative prospect theory [86] is regarded as one of the most prominent theories of individual choices under uncertainty. Cumulative prospect theory advances original rank-dependent models by considering *loss aversion* and *reference dependence*. Given the discrepancy between cumulative prospect theory and risk neutrality, it is expected that agents who behave according to cumulative prospect theory misreport their true beliefs under a proper scoring rule since they are no longer expected value maximizers.

We first show that, in theory, one should be able to obtain an agent's true belief from his misreported belief by solving a system of equations when the components that drive the agent's attitude towards uncertainty are known. In practice, however, there may be multiple true beliefs associated with the same reported belief. This happens because of the natural incompatibility between proper scoring rules and cumulative prospect theory, *i.e.*, traditional proper scoring rules do not satisfy an essential property called *comonotonicity* [45]. We provide a new explanation for why comonotonicity is a sufficient condition for making proper scoring rules indeed proper under cumulative prospect theory. Further, we suggest how to construct a comonotonic proper scoring rule from any proper scoring rule.

We also propose a new approach to obtain the true belief of an agent who takes decisions according to cumulative prospect theory when the components that drive the agent's attitude towards uncertainty are unknown. Our approach uses non-deterministic payments based on proper scoring rules.

4.1 Related Work

A standard assumption when using proper scoring rules is that the underlying agents are risk neutral. Such an assumption is often unrealistic when agents are human beings [80]. Focusing on the quadratic scoring rule, Winkler and Murphy [95] investigated the effects of nonlinear utility functions on how agents report their beliefs. For some specific utility functions, they showed that a risk-seeking agent reports a very sharp (narrow) probability distribution, whereas a risk-averse agent reports a distribution close to the

[2] That is, when agents rank the payoffs when calculating the utility of a decision.

uniform distribution. In other words, different utility functions induce different reporting behavior.

Winkler and Murphy's analysis assumed that agents are expected utility maximizers, where different risk attitudes are driven exclusively by utility functions. More modern models of decisions under uncertainty based on rank-dependent utilities assert that besides nonlinear utilities, probability sensitivity also plays a role in defining an agent's attitude towards uncertainty [33, 66, 73].

Focusing on binary outcomes, Offerman *et al.* [61] discussed how to calibrate *a posteriori* beliefs reported under the quadratic scoring rule by agents who take decisions based on rank-dependent utilities. Our work generalize the results by Offerman *et al.* by considering cumulative prospect theory [86], a decision theory based on rank-dependent utilities which also incorporates loss aversion and reference dependence. In other words, we discuss how to obtain the true belief of a cumulative prospect theory agent from his misreported belief under any proper scoring rule and for any number of outcomes.

Kothiyal *et al.* [45] noted that traditional proper scoring rules might not actually be proper under rank-dependent models. The reason for this is that traditional proper scoring rules do not satisfy comonotonicity, which implies that a reported belief might relate to many true beliefs and, thus, it cannot be used to uniquely determine the correct true belief. Focusing on binary outcomes, Kothiyal *et al.* [45] showed how to construct a comonotonic proper scoring rule from a positive proper scoring rule. We extend the results by Kothiyal *et al.* in two different ways. First, we provide a novel explanation for why comonotonicity is a sufficient condition for mapping each reported belief to a single true belief, thus allowing any true belief to be obtained from a reported belief. Second, we show how to construct a comonotonic proper scoring rule from any proper scoring rule and for any number of outcomes.

In all previously mentioned works, it was assumed that all the components that drive an agent's attitude towards uncertainty are known *a priori*, *e.g.*, they were previously elicited using methods such as the trade-off method [1, 90]. In this chapter, we also propose an approach to obtain an agent's true belief when those components are unknown by making probabilistic payments based on proper scoring rules. Different non-deterministic payment structures based on proper scoring rules have been proposed in the literature [4, 43]. Our payment structure is based on the work by Sandroni and Shmaya [71].

4.2 Proper Scoring Rules and Risk-Neutral Behavior

A basic assumption underlying proper scoring rules is that agents are *risk neutral*, *i.e.*, they behave so as to maximize their expected scores in (2.1). Consequently, a risk-neutral agent always honestly reports his belief under a strictly proper scoring rule R because $\arg\max_{\mathbf{q}} \mathbb{E}_{\mathbf{p}}\left[R(\mathbf{q}, \cdot)\right] = \mathbf{p}$.

The assumption of risk-neutral behavior is hardly compelling when the stakes are high [80], *e.g.*, when scores from a proper scoring rule are associated with high monetary values. It might be argued, however, that risk neutrality holds true when the stakes are low. For example, in his seminal work about the theoretical foundations of proper scoring rules, Savage [72] said the following regarding risk neutrality:

> "This assumption is not altogether unobjectionable; for it may imply that the person's utility function is linear in money. But such linearity assumptions are made almost throughout the present paper and are presumably tolerable if only moderate sums of money are involved." [72, p. 791]

In order to investigate whether or not risk-neutral behavior holds true when the scores from a proper scoring rule are associated with low monetary values, we performed an experiment on *Amazon's Mechanical Turk* (AMT), the online marketplace for work described in Section 2.2. Specifically, we designed a task on AMT that required participants to predict the outcomes of the 2013 NBA playoff games.

The participants were all older than 18 years old and resided in the United States of America. For each one of the 85 playoff games, we informed the participants about the names of the teams playing the game and the date of the game. Thereafter, we asked the participants to predict the result of each game twice. The first time, we informed the participants that besides the baseline payment of \$0.10 for completing the task, each participant could earn an additional bonus of up to \$0.10 based solely on his performance. To compute the bonus, we used a positive affine transformation of the quadratic scoring rule:

$$R(\mathbf{q}, \theta_x) = 5 \times \left(2q_x - \sum_{k=1}^{n} q_k^2\right) + 5$$

The resulting strictly proper scoring rule is positive, and its range is $[0, 10]$. After reporting the first prediction, we asked each participant to predict the result of the same game once more, but this time we informed the participant that he had already received a

bonus of \$0.10 cents. Furthermore, such a bonus would decrease according to his performance. In order to calculate the final bonus, we used a positive affine transformation of the logarithmic scoring rule:

$$R(\mathbf{q}, \theta_x) = \log q_x + 10$$

Since the logarithmic function is not well-defined when $q_x = 0$, we restricted the minimum probability value reported by a participant to 0.0000453, which implies that $\min R(\mathbf{q}, \theta_x) = \log 0.0000453 + 10 \approx 0$. Hence, the range of the final bonus was $[0, 10]$.

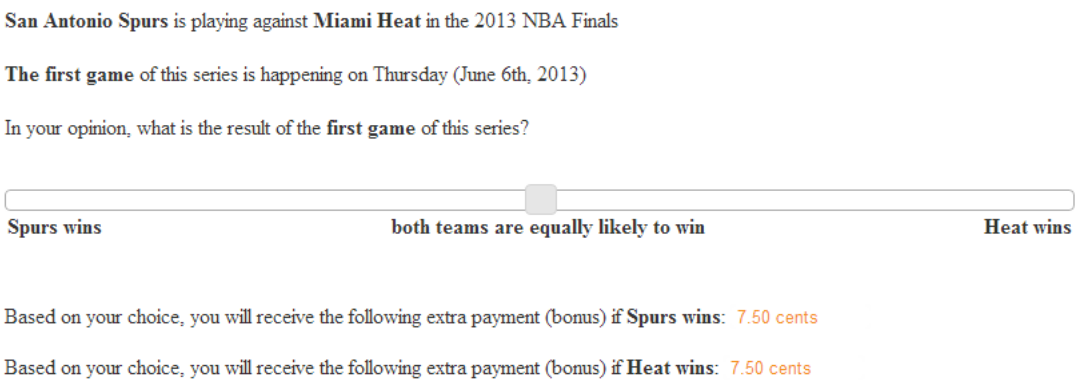
Instead of reporting probability values directly, the participants reported their beliefs by sliding a horizontal bar. The more a participant moved the bar towards one end, the higher the probability value associated with the underlying team would be. Figure 4.1 shows an example of the graphical interface used in our experiment. For each possible outcome, the participants could visualize the resulting bonus they would receive if the outcome turned out to be true.

We collected a total of 5822 predictions, *i.e.*, 2911 predictions under each scoring rule. If a participant is risk neutral, then his prediction of the result of a game does not change when different proper scoring rules are used[3]. In other words, the L_1 distance between the beliefs reported under two different proper scoring rules should be 0. Our experimental results show, however, that the mean of the L_1 distances between the reported beliefs across the 2911 data points is 0.34, the standard deviation being equal to 0.5. The mean distance is statistically significantly greater than 0 (p-value < 0.01, Wilcoxon rank-sum test), which implies that agents are not risk neutral. Thus, our experimental results show that different proper scoring rules induce different reporting behavior, even when the stakes are low.

We discuss later how reporting a belief under a proper scoring rule is equivalent to making a choice under uncertainty. Hence, the logical next step is to adapt proper scoring rules to a more appropriate model of decision-making under uncertainty. In this chapter, we adapt proper scoring rules to *Cumulative Prospect Theory* [86], which has been shown to provide a more accurate description of individual choice under uncertainty than many other decision theories, including expected utility theory and, consequently, risk neutrality.

[3]We note that the scoring rules used to compute the bonuses in our experiment are proper because a positive affine transformation of a proper scoring rule is still proper (see Proposition 1).

Figure 4.1: Example of the graphical interface used in the experiments.



4.3 Cumulative Prospect Theory

It has been reported that human beings are not risk neutral when the stakes are high [80]. In the previous section, we showed that the assumption of risk-neutral behavior is not empirically plausible when the stakes are low. Different than risk neutrality and, more broadly, expected utility theory, modern theories of individual choices under risk and uncertainty based on *ranks* assert that both sensitivity to payoffs and sensitivity to probabilities generate deviations from risk-neutral behavior. *Cumulative Prospect Theory* (CPT) [86] is one of such theory based on ranks which also incorporates loss aversion and reference dependence.

CPT is defined is terms of *prospects*, which are event-contingent payoffs. We use the notation $\mathbf{f} = [s_1 : \theta_1, \ldots, s_n : \theta_n]$ to denote a prospect $\mathbf{f}$ which yields a payoff of $s_k \in \Re$ if outcome θ_k occurs, for $k \in \{1, \ldots, n\}$. Since one can always rearrange outcomes and collapse similar payoffs, we assume without loss of generality that $s_n > s_{n-1} > \cdots > s_1$. The uncertainties regarding the occurrence of the outcomes are quantified in terms of a probability vector $\mathbf{p} = (p_1, \ldots, p_n)$. Consequently, we can represent a prospect as $\mathbf{f} = [s_1 : p_1, \ldots, s_n : p_n]$, which yields a payoff of $s_k \in \Re$ with probability p_k.

A prospect is called *positive* if all payoffs are nonnegative, *i.e.*, $s_n > \cdots > s_1 \geq 0$. We denote a positive prospect by $\mathbf{f}^+$. A prospect is called *negative* if all payoffs are nonpositive, *i.e.*, $0 \geq s_n > \cdots > s_1$. We refer to a negative prospect as $\mathbf{f}^-$. Finally, a *mixed* prospect $\mathbf{f}^\pm$ contains both positive and negative payoffs, *i.e.*, $s_n > \cdots > s_i \geq 0 > s_{i-1} > \cdots > s_1$.

In the expected utility theory, which subsumes risk neutrality, the utility of a prospect is the sum of the utilities of the payoffs, each weighted by its underlying probability. CPT

proposes two major modifications of the expected utility theory: 1) the utility of each payoff is multiplied by a *decision weight*, not by an additive probability; and 2) the carriers of value are gains and losses relative to a reference point, not final payoffs. Formally, the value an agent derives from a prospect is defined in terms of a *utility function*, a *loss-aversion parameter*, and cumulative *weighting functions*. We discuss these components in the following subsections before going into details about CPT utilities.

4.3.1 Utility Functions and the Loss-Aversion Parameter

Cumulative prospect theory asserts that the carriers of value are gains and losses relative to a reference point. We implicitly used the reference point equal to 0 in our definitions of positive, negative, and mixed prospects. Agents have different risk attitudes towards gains and losses, *i.e.*, with respect to payoffs above and below the reference point. Formally, the intrinsic value of a payoff s_k is defined in terms of a strictly increasing *value function* $V\colon \Re \to \Re$ satisfying:

$$V(s_k) = U(s_k) \text{ for } s_k \geq 0$$
$$V(s_k) = \lambda U(s_k) \text{ for } s_k < 0$$

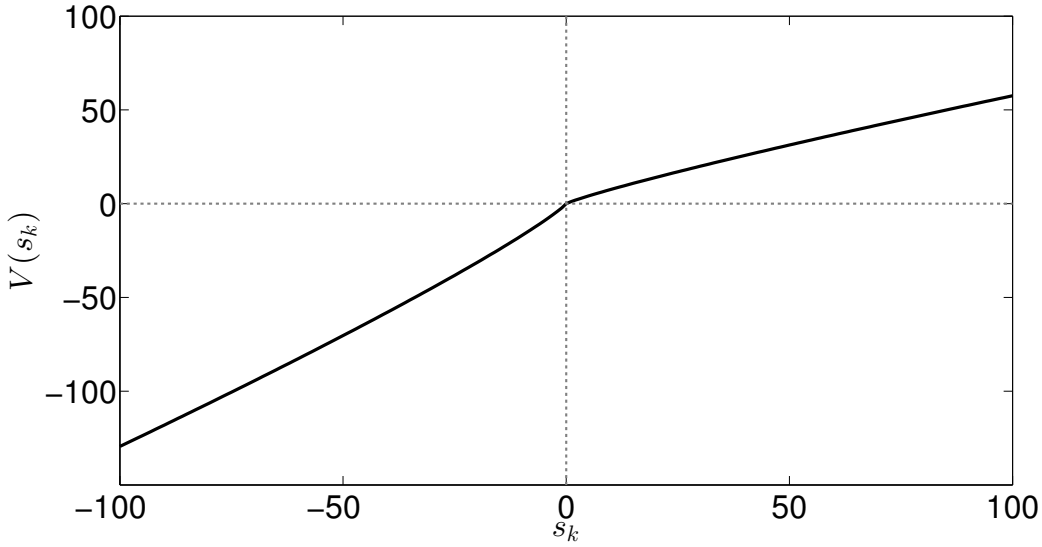
where $U\colon \Re \to \Re$ is a continuously differentiable strictly increasing *utility function* satisfying $U(s_k) \geq 0$, for $s_k \geq 0$, and $U(s_k) < 0$, for $s_k < 0$. The parameter $\lambda > 0$ is the *loss-aversion* parameter. The loss-aversion parameter captures the psychological phenomenon that losses usually loom larger than gains. As a consequence, the value function is steeper for losses than for gains, *i.e.*, $V'(s_k) < V'(-s_k)$, for $s_k \geq 0$. It is often the case that the parameter λ is defined in terms of the utility function U. For example, Tversky and Kahneman [86] implicitly assumed that $\lambda = -U(-1)$.

Another property of the value function V is that it is concave for gains, convex for losses, and it satisfies $V(0) = 0$. Tversky and Kahneman [86] proposed the following value function:

$$V(s_k) = \begin{cases} s_k^{\alpha} & \text{for } s_k \geq 0, \\ \lambda \cdot \left(-(-s_k)^{\beta}\right) & \text{otherwise} \end{cases} \tag{4.1}$$

where α and β are parameters of the power utility function:

Figure 4.2: Tversky and Kahneman's value function for $\alpha = \beta = 0.88, \lambda = 2.25$, and the reference point equal to 0.



$$U(s_k) = \begin{cases} s_k^{\alpha} & \text{for } s_k \geq 0, \\ \left(-(-s_k)^{\beta}\right) & \text{otherwise} \end{cases}$$

Figure 4.2 illustrates the value function in (4.1) for parameter values empirically determined by Tversky and Kahneman [86].

4.3.2 Weighting Functions

Cumulative prospect theory asserts that the weight associated with the value $V(s_k)$ is the difference between two transformed *ranks*, instead of an individual probability value p_k as in expected utility theory. For a positive prospect $\mathbf{f}^+$, the rank of a payoff s_k is the probability of $\mathbf{f}^+$ yielding a payoff weakly better than s_k, *i.e.*, the rank of s_k is equal to $p_{k+1}+p_{k+2}+\cdots+p_n$. The weight of $V(s_k)$ is then the marginal contribution of the individual probability p_k to its rank, *i.e.*, to the total probability of receiving better payoffs, measured in terms of a strictly increasing function $W^+\colon [0,1] \to [0,1]$. Formally, the weight of $V(s_k)$ is π_k^+, where:

$$\pi_n^+ = W^+(p_n) \tag{4.2}$$
$$\pi_k^+ = W^+\left(\sum_{x=k}^{n} p_x\right) - W^+\left(\sum_{x=k+1}^{n} p_x\right), \text{ for } k \in \{1, \ldots, n-1\}$$

For a negative prospect, the weight associated with $V(s_k)$, π_k^-, is the marginal contribution of the individual probability p_k to the total probability of receiving worse payoffs, measured in terms of a strictly increasing function $W^-\colon [0,1] \to [0,1]$, *i.e.*:

$$\pi_1^- = W^-(p_1) \tag{4.3}$$
$$\pi_k^- = W^-\left(\sum_{x=1}^{k} p_x\right) - W^-\left(\sum_{x=1}^{k-1} p_x\right), \text{ for } k \in \{2, \ldots, n\}$$

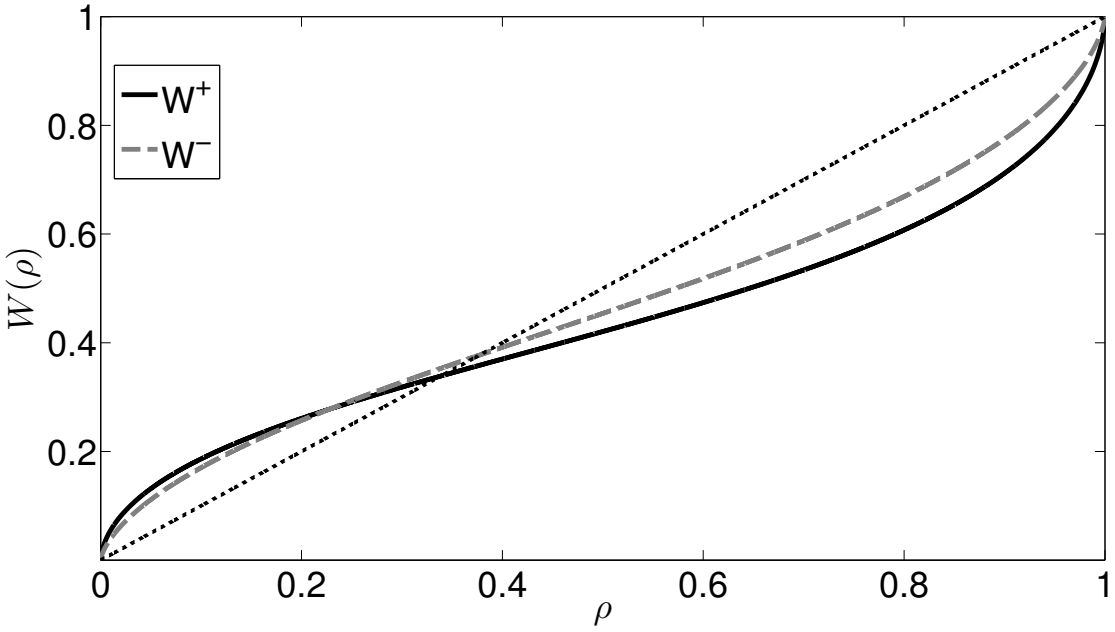
For a mixed prospect where $s_n > \cdots > s_i \geq 0 > s_{i-1} > \cdots > s_1$, we have that the weight associated with $V(s_k)$, for $k \in \{i, \ldots, n\}$, is π_k^+, whereas π_k^- is the weight associated with $V(s_k)$, for $k \in \{1, \ldots, i-1\}$. The π-values in (4.2) and (4.3) are traditionally referred to as *decision weights*. Decision weights reflect a cognitive bias concerning how human beings deal with probabilities and, thus, they should not be taken as a measure of an agent's true belief. Importantly, $(\pi_1^+, \ldots, \pi_n^+)$ and $(\pi_1^-, \ldots, \pi_n^-)$ are probability vectors, whereas the same is only true for $(\pi_1^-, \ldots, \pi_{i-1}^-, \pi_i^+, \ldots, \pi_n^+)$ when $W^-(\rho) + W^+(1-\rho) = 1$, for all $\rho \in [0,1]$.

The functionals W^+ and W^- are known as *weighting functions*. Common findings suggest that a weighting function is a nonlinear transformation of the probability scale that overweights small probabilities and underweights moderate and high probabilities [1, 86]. In other words, the weighting function displays an inverse-S shape: it is concave near 0 and convex near 1. Moreover, $W^+(0) = W^-(0) = 0$, and $W^+\left(\sum_{k=1}^{n} p_k\right) = W^-\left(\sum_{k=1}^{n} p_k\right) = 1$. The weighting functions proposed by Tversky and Kahneman [86] for gains and loses are:

$$W^+(\rho) = \frac{\rho^\gamma}{(\rho^\gamma + (1-\rho)^\gamma)^{\frac{1}{\gamma}}} \quad \text{and} \quad W^-(\rho) = \frac{\rho^\delta}{(\rho^\delta + (1-\rho)^\delta)^{\frac{1}{\delta}}} \tag{4.4}$$

where $\gamma, \delta \geq 0.28$ in order for the weighting functions to be strictly increasing. Figure 4.3 illustrates the above weighting functions for $\gamma = 0.61$ and $\delta = 0.69$, the parameter values empirically found by Tversky and Kahneman [86].

Figure 4.3: Tversky and Kahneman's weighting functions for $\gamma = 0.61$ and $\delta = 0.69$.



4.3.3 CPT Utility

Cumulative prospect theory asserts that risk attitudes are jointly determined by the weighting functions and the value function. Together, they result in a fourfold pattern of risk attitudes: risk aversion for gains and risk seeking for losses of high probability; risk seeking for gains and risk aversion for losses of low probability [86]. Let CPT($\mathbf{f}$) be the value an agent derives from a prospect according to cumulative prospect theory. For a positive prospect $\mathbf{f}^+ = [s_1 \colon p_1, \ldots, s_n \colon p_n]$, CPT($\mathbf{f}^+$) is defined as:

$$\text{CPT}(\mathbf{f}^+) = \sum_{k=1}^{n} \pi_k^+ V(s_k) \tag{4.5}$$

The CPT utility for a negative prospect $\mathbf{f}^- = [s_1 \colon p_1, \ldots, s_n \colon p_n]$ is:

$$\text{CPT}(\mathbf{f}^-) = \sum_{k=1}^{n} \pi_k^- V(s_k) \tag{4.6}$$

Finally, the CPT utility for a mixed prospect $\mathbf{f}^{\pm} = [s_1 \colon p_1, \ldots, s_n \colon p_n]$, where $s_n > \cdots > s_i \geq 0 > s_{i-1} > \cdots > s_1$, is:

$$\text{CPT}\left(\mathbf{f}^{\pm}\right) = \sum_{k=1}^{i-1} \pi_k^- V(s_k) + \sum_{k=i}^{n} \pi_k^+ V(s_k) \tag{4.7}$$

4.4 Proper Scoring Rules and Cumulative Prospect Theory

For ease of exposition and mathematical notation, consider a permutation of the outcomes for a given $\mathbf{q}$ so that the scores from a proper scoring rule R are ordered, *i.e.*, $R(\mathbf{q}, \theta_n) > R(\mathbf{q}, \theta_{n-1}) > \cdots > R(\mathbf{q}, \theta_1)$. We note that the scores resulting from a proper scoring rule R can be stated in terms of a prospect, *i.e.*, $[R(\mathbf{q}, \theta_1) \colon p_1, \ldots, R(\mathbf{q}, \theta_n) \colon p_n]$. In other words, when reporting a belief $\mathbf{q}$, an agent is essentially defining the payoffs of a prospect, where the associated probabilities are subjective probabilities. Reporting a belief $\mathbf{q}$ is then equivalent to taking a decision under uncertainty by choosing a prospect amongst a potentially infinite number of prospects. Consequently, an agent's reporting behavior can be analyzed from the perspective of decision theories such as cumulative prospect theory.

In general, the utility derived by an agent who behaves according to cumulative prospect theory from a given prospect is only equal to the utility derived by a risk-neutral agent when: 1) the weighting functions are identity functions; 2) the utility function is a linear function; and 3) the loss aversion parameter is equal to 1. Hence, cumulative prospect theory subsumes risk neutrality and, more broadly, expected utility theory. As discussed before, common findings suggest that the weighting functions are nonlinear functions, the utility function is concave for gains and convex for losses, and the loss aversion parameter is greater than 1. Thus, risk-neutral agents and agents with CPT utilities are expected to value prospects in different ways and, consequently, to behave differently when reporting their beliefs under a proper scoring rule.

In the following subsections, we characterize how experts with CPT utilities report their beliefs. Formally, we obtain systems of equations for different proper scoring rules that describe how agents with CPT utilities report their beliefs. From the characterization, we then describe how to obtain their true beliefs, *i.e.*, we discuss how to solve those systems of equations in order to obtain an agent's true belief from his misreported belief. Naturally, we start by assuming that an agent reports his belief so as to maximize his CPT utility. In other words, each agent is solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{q}}{\text{maximize}} \quad & \text{CPT}(\mathbf{q}) \\ \text{subject to} \quad & \sum_{k=1}^{n} q_k = 1 \\ & q_k \geq 0, k \in \{1, \ldots, n\} \end{aligned}$$

We abuse notation by dropping the underlying scoring rule R and by writing $\text{CPT}(\mathbf{q})$, instead of $\text{CPT}(\mathbf{f})$, for $\mathbf{f} = [R(\mathbf{q}, \theta_1) \colon p_1, \ldots, R(\mathbf{q}, \theta_n) \colon p_n]$. Given that the CPT functional is continuous and that the domain of every q_k is compact, an optimal $\mathbf{q}$ always exists. There may exist several optima, in which case one optimum is arbitrarily selected to be the reported probability vector $\mathbf{q}$. Since all the constraints are linear, it must be the case that the reported probability vector $\mathbf{q}$ satisfies the Karush-Kuhn-Tucker (KKT) conditions. The Lagrangian associated with the above problem is:

$$L(\mathbf{q}, \tau, \boldsymbol{\mu}) = \text{CPT}(\mathbf{q}) - \tau \left(\sum_{k=1}^{n} q_k - 1 \right) + \sum_{k=1}^{n} \mu_k q_k \tag{4.8}$$

where the dual feasibility and the complementary slackness are:

$$\begin{aligned} \mu_k \geq 0, \quad & \text{for } k \in \{1, \ldots, n\} \\ \mu_k q_k = 0, \quad & \text{for } k \in \{1, \ldots, n\} \end{aligned}$$

Let $\mathbf{q}^* = (q_1^*, \ldots, q_n^*), \tau^*, \boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)$ be the optimal points for the Lagrangian. Henceforth, we assume that the constraint $q_k \geq 0$ is not binding at the optimal, *i.e.*, $q_k^* > 0$, for $k \in \{1, \ldots, n\}$. Consequently, $\mu_k^* = 0$, for $k \in \{1, \ldots, n\}$. This assumption allows us to remove $\boldsymbol{\mu}$ from the Lagrangian. We note that without such an assumption, we might run into technical problems whenever the underlying scoring rule is not bounded, *e.g.*, the resulting CPT utility is equal $-\infty$ when using the logarithmic scoring rule and $q_k^* = 0$, for some $k \in \{1, \ldots, n\}$. However, we allow q_k^* to be arbitrarily close to 0.

In the following subsections, we discuss the implications of the above optimization problem for, respectively, positive, negative, and mixed proper scoring rules. That is, we derive systems of equations in terms of an agent's value function, weighting functions, true belief, and reported belief that, by construction, must be valid. In the next section, we discuss how to solve these systems of equations to obtain an agent's true belief from his misreported belief when the agent's value function and weighting functions are known.

4.4.1 Positive Scoring Rules

Suppose that an agent's belief is elicited through a positive scoring rule R, for $R(\mathbf{q}, \theta_n) > R(\mathbf{q}, \theta_{n-1}) > \cdots > R(\mathbf{q}, \theta_1) \geq 0$. The CPT utility of that agent is then:

$$\text{CPT}(\mathbf{q}) = \sum_{k=1}^{n} \pi_k^+ V(R(\mathbf{q}, \theta_k))$$

Consider the Lagrangian in (4.8). Given that $\mathbf{q}^* = (q_1^*, \ldots, q_n^*)$, τ^*, and $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)$ are optimal points, we obtain the KKT conditions:

$$\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_k^*} = 0, \text{ for } k \in \{1, \ldots, n\}$$

For illustration's sake, consider the partial derivative of L with respect to q_n^*:

$$\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_n^*} = \left(\sum_{k=1}^{n} \pi_k^+ \frac{\partial V(R(\mathbf{q}^*, \theta_k))}{\partial q_n^*} \right) - \tau^* = 0$$

By considering all partial derivatives $\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_k^*}$, for $k \in \{1, \ldots, n\}$, we end up with the following system of equations:

$$\begin{aligned}
\pi_1^+ \frac{\partial V(R(\mathbf{q}^*, \theta_1))}{\partial q_1^*} + \pi_2^+ \frac{\partial V(R(\mathbf{q}^*, \theta_2))}{\partial q_1^*} + \cdots + \pi_n^+ \frac{\partial V(R(\mathbf{q}^*, \theta_n))}{\partial q_1^*} &= \tau^* \\
\pi_1^+ \frac{\partial V(R(\mathbf{q}^*, \theta_1))}{\partial q_2^*} + \pi_2^+ \frac{\partial V(R(\mathbf{q}^*, \theta_2))}{\partial q_2^*} + \cdots + \pi_n^+ \frac{\partial V(R(\mathbf{q}^*, \theta_n))}{\partial q_2^*} &= \tau^* \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad & \vdots \\
\pi_1^+ \frac{\partial V(R(\mathbf{q}^*, \theta_1))}{\partial q_n^*} + \pi_2^+ \frac{\partial V(R(\mathbf{q}^*, \theta_2))}{\partial q_n^*} + \cdots + \pi_n^+ \frac{\partial V(R(\mathbf{q}^*, \theta_n))}{\partial q_n^*} &= \tau^*
\end{aligned} \tag{4.9}$$

We discuss in Section 4.5.1 how to solve the above system of equations.

4.4.2 Negative Scoring Rules

Now, suppose that an agent's belief is elicited through a negative scoring rule R, for $0 \geq R(\mathbf{q}, \theta_n) > R(\mathbf{q}, \theta_{n-1}) > \cdots > R(\mathbf{q}, \theta_1)$. The CPT utility of that agent is:

$$\text{CPT}(\mathbf{q}) = \sum_{k=1}^{n} \pi_k^- V(R(\mathbf{q}, \theta_k))$$

By considering all partial derivatives $\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_k^*}$, for $k \in \{1, \ldots, n\}$, we end up with the following system of equations:

$$\begin{aligned}
\pi_1^- \frac{\partial V(R(\mathbf{q}^*, \theta_1))}{\partial q_1^*} + \pi_2^- \frac{\partial V(R(\mathbf{q}^*, \theta_2))}{\partial q_1^*} + \cdots + \pi_n^- \frac{\partial V(R(\mathbf{q}^*, \theta_n))}{\partial q_1^*} &= \tau^* \\
\pi_1^- \frac{\partial V(R(\mathbf{q}^*, \theta_1))}{\partial q_2^*} + \pi_2^- \frac{\partial V(R(\mathbf{q}^*, \theta_2))}{\partial q_2^*} + \cdots + \pi_n^- \frac{\partial V(R(\mathbf{q}^*, \theta_n))}{\partial q_2^*} &= \tau^* \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \\
\pi_1^- \frac{\partial V(R(\mathbf{q}^*, \theta_1))}{\partial q_n^*} + \pi_2^- \frac{\partial V(R(\mathbf{q}^*, \theta_2))}{\partial q_n^*} + \cdots + \pi_n^- \frac{\partial V(R(\mathbf{q}^*, \theta_n))}{\partial q_n^*} &= \tau^*
\end{aligned} \tag{4.10}$$

We discuss in Section 4.5.2 how to solve the above system of equations.

4.4.3 Mixed Scoring Rules

Finally, suppose that an agent's belief is elicited through a mixed scoring rule R, for $R(\mathbf{q}, \theta_n) > \cdots > R(\mathbf{q}, \theta_i) \geq 0 > R(\mathbf{q}, \theta_{i-1}) > \cdots > R(\mathbf{q}, \theta_1)$. The CPT utility of that agent is then:

$$\text{CPT}(\mathbf{q}) = \sum_{k=1}^{i-1} \pi_k^- V(R(\mathbf{q}, \theta_k)) + \sum_{k=i}^{n} \pi_k^+ V(R(\mathbf{q}, \theta_k))$$

By considering all partial derivatives $\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_k^*}$, for $k \in \{1, \ldots, n\}$, we end up with the following system of equations:

$$\begin{aligned}
&\sum_{k=1}^{i-1} \pi_k^- \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_1^*} + \sum_{k=i}^{n} \pi_k^+ \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_1^*} = \tau^* \\
&\sum_{k=1}^{i-1} \pi_k^- \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_2^*} + \sum_{k=i}^{n} \pi_k^+ \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_2^*} = \tau^* \\
&\quad \vdots \qquad\qquad\qquad\qquad \vdots \qquad\quad \vdots \\
&\sum_{k=1}^{i-1} \pi_k^- \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_n^*} + \sum_{k=i}^{n} \pi_k^+ \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_n^*} = \tau^*
\end{aligned} \tag{4.11}$$

We discuss in Section 4.5.3 how to solve the above system of equations.

4.5 Obtaining True Beliefs from Misreported Beliefs When the Value Function and Weighting Functions Are Known

In this section, we discuss how to solve the previous systems of equations to obtain an agent's true belief from his misreported belief under a traditional proper scoring rule R. In our analysis, we implicitly assume that the components of the CPT utility are known, *i.e.*, the utility function, the loss-aversion parameter, and the weighting functions were previously elicited through some elicitation procedure (*e.g.*, [1, 2, 90]). This scenario was previously studied by Offerman *et al.* [61]. However, Offerman *et al.* focused only on the positive version of the quadratic scoring rule for binary outcomes ($n = 2$). In the following subsections, we generalize the results by Offerman *et al.* by considering any finite number of outcomes as well as positive, negative, and mixed proper scoring rules.

4.5.1 Positive Scoring Rules

When the value function V is known, the terms $\frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_x^*}$ in (4.9), for all $k, x \in \{1, \ldots, n\}$, become known numerical values. Given the definition of π_k^+ in (4.2), the system of equations in (4.9) can then be written as:

$$\begin{bmatrix} -1 & a_{1,2} & \cdots & a_{1,n} \\ -1 & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -1 & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} \tau^* \\ W^+\left(\sum_{k=2}^{n} p_k\right) \\ \vdots \\ W^+(p_{n-1}+p_n) \\ W^+(p_n) \end{bmatrix} = \begin{bmatrix} -\frac{\partial V(R(\mathbf{q},\theta_1))}{\partial q_1^*} \\ -\frac{\partial V(R(\mathbf{q},\theta_1))}{\partial q_2^*} \\ \vdots \\ -\frac{\partial V(R(\mathbf{q},\theta_1))}{\partial q_n^*} \end{bmatrix} \tag{4.12}$$

where $a_{x,k} = \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_x^*} - \frac{\partial V(R(\mathbf{q},\theta_{k-1}))}{\partial q_x^*}$. Hence, the system of equations in (4.9) becomes a system of linear equations when one considers all the $W^+(\cdot)$ as well as τ^* as variables. Let $W^+(p_n) = y_n$ after solving the above system of linear equations. Since W^+ is strictly increasing, it admits an inverse function. Consequently, $p_n = W^{+^{-1}}(y_n)$. Similarly, let $W^+(p_n + p_{n-1}) = y_{n-1} \implies p_{n-1} = W^{+^{-1}}(y_{n-1}) - p_n = W^{+^{-1}}(y_{n-1}) - W^{+^{-1}}(y_n)$. More generally, for all $k \in \{2, \ldots, n-1\}$, we obtain p_k by using backward substitution, *i.e.*, by solving the equation $p_k = W^{+^{-1}}(y_k) - W^{+^{-1}}(y_{k+1})$. Lastly, $p_1 = 1 - \sum_{k=2}^{n} p_k$.

Numerical Example

Consider three exhaustive and mutually exclusive outcomes ($n = 3$) and let $R(\mathbf{q}, \theta_x) = 2q_x - \left(\sum_{k=1}^{3} q_k^2\right) + 1$, *i.e.*, R is a positive affine transformation of the quadratic scoring rule with range $[0, 2]$. Suppose that an agent reports the belief $\mathbf{q}^* = (0.1, 0.3, 0.6)$. Consequently, $R(\mathbf{q}^*, \theta_3) > R(\mathbf{q}^*, \theta_2) > R(\mathbf{q}^*, \theta_1) > 0$. For illustration's sake, consider the last equation in (4.9):

$$\left[1 - W^+(p_2 + p_3)\right] \frac{\partial V(R(\mathbf{q}, \theta_1))}{\partial q_3^*} + \left[W^+(p_2 + p_3) - W^+(p_3)\right] \frac{\partial V(R(\mathbf{q}, \theta_2))}{\partial q_3^*} + W^+(p_3) \frac{\partial V(R(\mathbf{q}, \theta_3))}{\partial q_3^*} = \tau^*$$

Let V be the value function in (4.1) with parameter $\alpha = 0.88$. Consequently, the above equation can be written as follows:

$$1.6983 \times W^+(p_3) + 0.0553 \times W^+(p_2 + p_3) - \tau^* = 1.0949$$

Repeating the same procedure for all partial derivatives $\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_k^*}$, for $k \in \{1, 2, 3\}$, we end up with the following system of linear equations:

$$
\begin{aligned}
0.0086 \times W^+(p_3) - 1.8155 \times W^+(p_2+p_3) - \tau^* &= -1.6423 \\
-1.7068 \times W^+(p_3) + 1.7602 \times W^+(p_2+p_3) - \tau^* &= 0.5474 \\
1.6983 \times W^+(p_3) + 0.0553 \times W^+(p_2+p_3) - \tau^* &= 1.0949
\end{aligned}
$$

The above system of linear equations is equivalent to the system in (4.12). After solving the above system of equations, we obtain: $W^+(p_3) = 0.6151$, $W^+(p_2+p_3) = 0.9075$, and $\tau^* = 4 \times 10^{-11}$. Let W^+ be the weighting function proposed by Tversky and Kahneman [86] shown in (4.4) with parameter $\gamma = 0.61$. We then obtain $p_3 = {W^+}^{-1}(0.6151) = 0.809$, $p_2 = {W^+}^{-1}(0.9075) - p_3 = 0.1802$, and $p_1 = 1 - p_3 - p_2 = 0.0108$. Thus, the agent's true belief is equal to $\mathbf{p} = (0.0108, 0.1802, 0.809)$.

4.5.2 Negative Scoring Rules

For a negative scoring rule R, and given the definition of π_k^- in (4.3), the system of equations in (4.10) can then be written as:

$$
\begin{bmatrix}
b_{1,1} & \cdots & b_{1,n-1} & -1 \\
b_{2,1} & \cdots & b_{2,n-1} & -1 \\
\vdots & \ddots & \vdots & \vdots \\
b_{n,1} & \cdots & b_{n,n-1} & -1
\end{bmatrix}
\begin{bmatrix}
W^-(p_1) \\
W^-(p_1+p_2) \\
\vdots \\
W^-\left(\sum_{k=1}^{n-1} p_k\right) \\
\tau^*
\end{bmatrix}
=
\begin{bmatrix}
-\frac{\partial V(R(\mathbf{q},\theta_n))}{\partial q_1^*} \\
-\frac{\partial V(R(\mathbf{q},\theta_n))}{\partial q_2^*} \\
\vdots \\
-\frac{\partial V(R(\mathbf{q},\theta_n))}{\partial q_n^*}
\end{bmatrix}
\tag{4.13}
$$

where $b_{x,k} = \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_x^*} - \frac{\partial V(R(\mathbf{q},\theta_{k+1}))}{\partial q_x^*}$. Hence, the system of equations in (4.10) becomes a system of linear equations when one considers all the $W^-(\cdot)$ as well as τ^* as variables. Let $W^-(p_1) = y_1$ after solving the above system of linear equations. Since W^- is strictly increasing, it admits an inverse function. Consequently, $p_1 = {W^-}^{-1}(y_1)$. Similarly, let $W^-(p_1+p_2) = y_2 \implies p_2 = {W^-}^{-1}(y_2) - p_1 = {W^-}^{-1}(y_2) - {W^-}^{-1}(y_1)$. More generally, for all $k \in \{2, \ldots, n-1\}$, we obtain p_k by using forward substitution, *i.e.*, by solving the equation $p_k = {W^-}^{-1}(y_k) - {W^-}^{-1}(y_{k-1})$. Lastly, $p_n = 1 - \sum_{k=1}^{n-1} p_k$.

Numerical Example

Consider three exhaustive and mutually exclusive outcomes ($n = 3$) and let $R(\mathbf{q}, \theta_x) = 2q_x - \left(\sum_{k=1}^{3} q_k^2\right) - 1$, *i.e.*, R is a positive affine transformation of the quadratic scoring

rule with range $[-2, 0]$. Suppose that an agent reports the belief $\mathbf{q}^* = (0.1, 0.3, 0.6)$. Consequently, $0 > R(\mathbf{q}^*, \theta_3) > R(\mathbf{q}^*, \theta_2) > R(\mathbf{q}^*, \theta_1)$. For illustration's sake, consider the last equation in (4.10):

$$W^-(p_1)\frac{\partial V(R(\mathbf{q},\theta_1))}{\partial q_3^*} + \left[W^-(p_1+p_2) - W^-(p_1)\right]\frac{\partial V(R(\mathbf{q},\theta_2))}{\partial q_3^*} + \left[1 - W^-(p_1+p_2)\right]\frac{\partial V(R(\mathbf{q},\theta_3))}{\partial q_3^*} = \tau^*$$

Let V be equal to the value function in (4.1) with parameters $\beta = 0.88$ and $\lambda = 2.25$. Consequently, the above equation can be written as follows:

$$0.1084 \times W^-(p_1) - 4.2813 \times W^-(p_1+p_2) - \tau^* = -1.8619$$

Repeating the same procedure for all partial derivatives $\frac{\partial L(\mathbf{q}^*,\tau^*,\boldsymbol{\mu}^*)}{\partial q_k^*}$, for $k \in \{1, 2, 3\}$, we end up with the following system of linear equations:

$$\begin{aligned}
3.8697 \times W^-(p_1) + 0.0622 \times W^-(p_1+p_2) - \tau^* &= 0.4655 \\
-3.9781 \times W^-(p_1) + 4.2191 \times W^-(p_1+p_2) - \tau^* &= 1.3964 \\
0.1084 \times W^-(p_1) - 4.2813 \times W^-(p_1+p_2) - \tau^* &= -1.8619
\end{aligned}$$

The above system of linear equations is equivalent to the system in (4.13). After solving the above system of equations, we obtain: $W^+(p_1) = 0.1132$, $W^+(p_1+p_2) = 0.4377$, and $\tau^* = 10^{-16}$. Let W^- be the weighting function proposed by Tversky and Kahneman [86] shown in (4.4) with parameter $\delta = 0.69$. We then obtain $p_1 = {W^-}^{-1}(0.1132) = 0.0513$, $p_2 = {W^-}^{-1}(0.4377) - p_1 = 0.4226$, and $p_3 = 1 - p_1 - p_2 = 0.5261$. Thus, the agent's true belief is equal to $\mathbf{p} = (0.0513, 0.4226, 0.5261)$.

4.5.3 Mixed Scoring Rules

For a mixed scoring rule R, the system of equations in (4.11) can then be written as:

$$\begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,i-1} & a_{1,i} & \cdots & a_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,i-1} & a_{2,i} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,i-1} & a_{n,i} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} W^-(p_1) \\ W^-(p_1+p_2) \\ \vdots \\ W^-\left(\sum_{k=1}^{i-1} p_k\right) \\ W^+\left(\sum_{k=i}^{n} p_k\right) \\ \vdots \\ W^+(p_n+p_{n-1}) \\ W^+(p_n) \end{bmatrix} = \begin{bmatrix} \tau^* \\ \tau^* \\ \vdots \\ \tau^* \end{bmatrix} \tag{4.14}$$

where $a_{x,k} = \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_x^*} - \frac{\partial V(R(\mathbf{q},\theta_{k-1}))}{\partial q_x^*}$, and $b_{x,k} = \frac{\partial V(R(\mathbf{q},\theta_k))}{\partial q_x^*} - \frac{\partial V(R(\mathbf{q},\theta_{k+1}))}{\partial q_x^*}$. By considering all the $W(\cdot)$ as well as τ^* as variables, the above system of equations becomes a system of linear equations with $n+1$ variables and n equations. The true belief $\mathbf{p} = (p_1, \ldots, p_n)$ can be obtained from (4.14) by following the procedure described below:

1. Solve (4.14) for $W^-\left(\sum_{k=1}^{i-1} p_k\right)$ in terms of τ^*, *i.e.*, $W^-\left(\sum_{k=1}^{i-1} p_k\right) = y_{i-1} \times \tau^* \implies \sum_{k=1}^{i-1} p_k = W^{-^{-1}}(y_{i-1} \times \tau^*)$, where $y_{i-1} \in \Re$ is a known numerical value;
2. Solve (4.14) for $W^+\left(\sum_{k=i}^{n} p_k\right)$ in terms of τ^*, *i.e.*, $W^+\left(\sum_{k=i}^{n} p_k\right) = y_i \times \tau^* \implies \sum_{k=i}^{n} p_k = W^{+^{-1}}(y_i \times \tau^*)$, where $y_i \in \Re$ is a known numerical value;
3. Combine the results from the first and second steps to obtain τ^*, *i.e.*, $\sum_{k=1}^{i-1} p_k + \sum_{k=i}^{n} p_k = W^{-^{-1}}(y_{i-1}\times\tau^*) + W^{+^{-1}}(y_i \times \tau^*) \implies 1 = W^{-^{-1}}(y_{i-1}\times\tau^*) + W^{+^{-1}}(y_i \times \tau^*)$.

Given that W^+, W^-, y_i, and y_{i-1} are known, the value of τ^* can be computed numerically after the third step. We discuss an approach to do so in the following subsection. Hence, the system of equations in (4.14) becomes a system of linear equations with n variables and n equations. The true belief $\mathbf{p}$ can then be obtained by following the extra steps described below:

4. Let $W^-(p_1) = y_1 \times \tau^*$. Consequently, $p_1 = W^{-^{-1}}(y_1 \times \tau^*)$;
5. Let $W^+(p_n) = y_n \times \tau^*$. Consequently, $p_n = W^{+^{-1}}(y_n \times \tau^*)$;
6. For all $k \in \{2, \ldots, i-1\}$, we obtain p_k as described in Section 4.5.2, *i.e.*, by solving the equation $p_k = W^{-^{-1}}(y_k \times \tau^*) - W^{-^{-1}}(y_{k-1} \times \tau^*)$;
7. For all $k \in \{i, \ldots, n-1\}$, we obtain p_k as described in Section 4.5.1, *i.e.*, by solving the equation $p_k = W^{+^{-1}}(y_k \times \tau^*) - W^{+^{-1}}(y_{k+1} \times \tau^*)$.

Numerical Example

Consider three exhaustive and mutually exclusive outcomes ($n = 3$) and let $R(\mathbf{q}, \theta_x) = 2q_x - \left(\sum_{k=1}^{3} q_k^2\right)$, *i.e.*, R is the quadratic scoring rule with range $[-1, 1]$. Suppose that an agent reports the belief $\mathbf{q}^* = (0.1, 0.3, 0.6)$. Consequently, $R(\mathbf{q}^*, \theta_3) > R(\mathbf{q}^*, \theta_2) > 0 > R(\mathbf{q}^*, \theta_1)$. For illustration's sake, consider the last equation in (4.11):

$$W^-(p_1)\frac{\partial V(R(\mathbf{q}, \theta_1))}{\partial q_3^*} + \left[W^+(p_2 + p_3) - W^+(p_3)\right]\frac{\partial V(R(\mathbf{q}, \theta_2))}{\partial q_3^*} + W^+(p_3)\frac{\partial V(R(\mathbf{q}, \theta_3))}{\partial q_3^*} = \tau^*$$

Let V be equal to the value function in (4.1) with parameters $\alpha = \beta = 0.88$ and $\lambda = 2.25$. Consequently, the above equation can be written as follows:

$$-2.7929 \times W^-(p_1) - 1.3370 \times W^+(p_2 + p_3) + 2.0669 \times W^+(p_3) = \tau^*$$

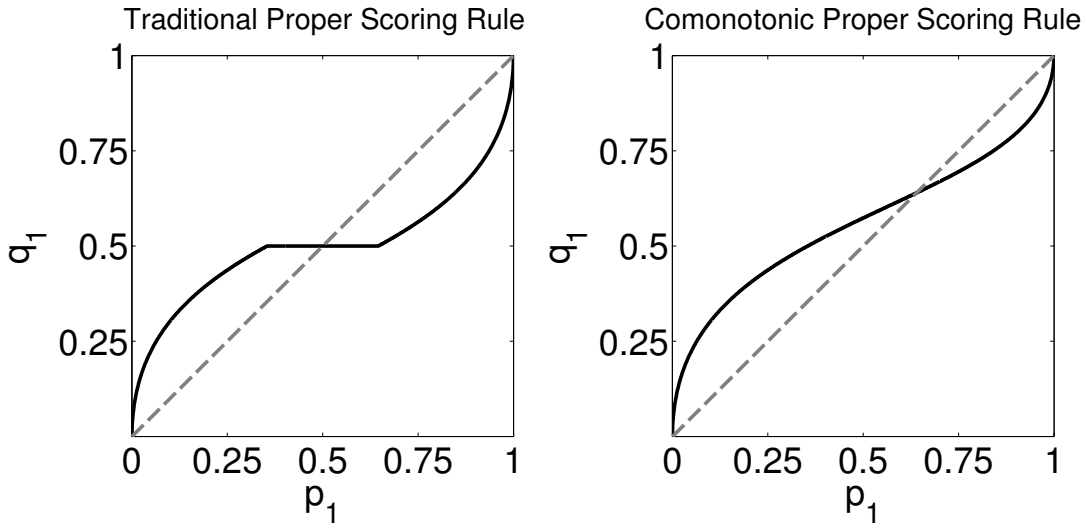
Repeating the same procedure for all partial derivatives $\frac{\partial L(\mathbf{q}^*, \tau^*, \boldsymbol{\mu}^*)}{\partial q_k^*}$, for $k \in \{1, 2, 3\}$, we end up with the following system of linear equations:

$$\begin{aligned}
4.1893 \times W^-(p_1) - 0.2228 \times W^+(p_2 + p_3) + 0.0404 \times W^+(p_3) &= \tau^* \\
-1.3964 \times W^-(p_1) + 1.5598 \times W^+(p_2 + p_3) - 2.1072 \times W^+(p_3) &= \tau^* \\
-2.7929 \times W^-(p_1) - 1.3370 \times W^+(p_2 + p_3) + 2.0669 \times W^+(p_3) &= \tau^*
\end{aligned}$$

The above system of linear equations is equivalent to the system in (4.14) for $i = 2$. After solving the above system of equations, we obtain: $W^-(p_1) = 1959.6882 \times \tau^*, W^+(p_2 + p_3) = 42283.3111 \times \tau^*$, and $W^+(p_3) = 30000 \times \tau^*$. Thus, $p_1 = W^{-^{-1}}(1959.6882 \times \tau^*), p_2 + p_3 = W^{+^{-1}}(42283.3111 \times \tau^*)$, and then $1 = W^{-^{-1}}(1959.6882 \times \tau^*) + W^{+^{-1}}(42283.3111 \times \tau^*)$.

For ease of exposition, let $\psi = 1959.6882 \times \tau^*$, and $\phi = 42283.3111 \times \tau^*$. By construction, $\psi, \phi \in [0, 1]$. Moreover, $\phi = \frac{42283.3111}{1959.6882} \times \psi$. Thus, to find the value of τ^*, we just need to find the value of $\psi \in [0, 1]$ such that $1 = W^{-^{-1}}(\psi) + W^{+^{-1}}\left(\frac{42283.3111}{1959.6882} \times \psi\right)$. Let W^+ and W^- be the weighting functions proposed by Tversky and Kahneman [86] shown in (4.4) with parameters $\gamma = 0.61$ and $\delta = 0.69$. We find numerically that $\phi = 0.0420$ and, consequently, $\tau^* = 0.00002143$. Finally, we obtain that $p_1 = W^{-^{-1}}(1959.6882 \times \tau^*) = 0.011$, $p_3 = W^{+^{-1}}(30000 \times \tau^*) = 0.8392$, and $p_2 = W^{+^{-1}}(42283.3111 \times \tau^*) - p_3 = 0.1498$. Thus, the agent's true belief is equal to $\mathbf{p} = (0.011, 0.1498, 0.8392)$.

Figure 4.4: (LEFT) The reported probability of 0.5 *relates to many degrees of true belief under a traditional proper scoring rule. (RIGHT) Each reported belief is mapped to a single true belief under a comonotonic proper scoring rule.*



4.6 Comonotonic Proper Scoring Rules

In the previous section, we showed how to obtain an agent's true belief from his reported belief by solving a system of equations. Unfortunately, the obtained belief is not always the agent's correct true belief. Under a proper scoring rule, agents with CPT utilities and different true beliefs might end up reporting the same belief. This point was noted by Offerman *et al.* [61] for the positive version of the quadratic scoring rule and binary outcomes. For illustration's sake, consider an agent with belief $\mathbf{p} = (p_1, p_2) = (p_1, 1 - p_1)$, who reports $\mathbf{q} = (q_1, q_2) = (q_1, 1-q_1)$ under the proper scoring rule considered by Offerman *et al.*, *i.e.*, $R(\mathbf{q}, \theta_x) = 2q_x + (\sum_{k=1}^{n} q_k^2) + 1$. The left part of Figure 4.4 illustrates how an agent with CPT utility reports q_1 in terms of p_1 when W^+ is equal to the weighting function in (4.4) with parameter $\gamma = 0.61$, and the value function is equal to the function in (4.1) with parameter $\alpha = 0.88$.

The most striking feature of the left part of Figure 4.4 is that the curve is flat around $p_1 = 0.5$. As explained by Offerman *et al.* [61], the risk aversion generated by the CPT utility is so strong for a subjective probability around 0.5 that an agent makes the safe choice of reporting 0.5. Thus, proper scoring rules may be insensitive to small changes in the neighborhood of 0.5 for binary outcomes. As a consequence, a reported probability of

0.5 relates to many degrees of true belief and, thus, it cannot be used to uniquely determine the correct true belief. As Kothiyal *et al.* [45] mentioned, if properness is to be taken to mean that all degrees of belief can be identified, then proper scoring rules are no longer proper under CPT.

The implication of the above problem in our work is that there are cases in which solutions for the systems of equations in (4.12), (4.13), and (4.14) might not actually be an agent's true belief. Kothiyal *et al.* [45] pointed out that such a problem is due to the fact that decision weights change in a drastic non-smooth manner when the ordering of the scores for any two outcomes changes. Mathematically, the abovementioned problem occurs when a proper scoring rule R allows $R(\mathbf{q}, \theta_x) = R(\mathbf{q}, \theta_y)$, for some probability vector $\mathbf{q}$, and $x, y \in \{1, \ldots, n\}$.

For ease of exposition and mathematical notation, we have assumed thus far that $R(\mathbf{q}, \theta_n) > R(\mathbf{q}, \theta_{n-1}) > \cdots > R(\mathbf{q}, \theta_1)$. We claimed that such an assumption was without loss of generality because the outcomes could always be rearranged *a posteriori*. In this section, however, we do not allow the outcomes to be rearranged. Moreover, we assume a strict ordering of the outcomes such that $R(\mathbf{q}, \theta_n) > R(\mathbf{q}, \theta_{n-1}) > \cdots > R(\mathbf{q}, \theta_1)$, *i.e.*, an agent will always receive the highest score when outcome θ_n happens, the second highest score when outcome θ_{n-1} happens, and so on, no matter what he reports. In other words, the scoring rule R satisfies a property called *comonotonicity* [45, 73].

Focusing on binary outcomes, Kothiyal *et al.* [45] suggested that one can avoid the non-smooth changes in decision weights, and the underlying analytical problems, when R is a comonotonic and positive proper scoring rule. In the following subsections, we provide a detailed explanation on why comonotonic proper scoring rules map each reported belief to a single true belief. We show that this happens not only for binary outcomes and positive proper scoring rules, but for any number of outcomes as well as positive, negative, and mixed proper scoring rules. In other words, we give a sufficient condition under which proper scoring rules are indeed proper under CPT.

4.6.1 Positive Scoring Rules

Consider a positive proper scoring rule R that satisfies comonotonicity, *i.e.*, $R(\mathbf{q}, \theta_n) > \cdots > R(\mathbf{q}, \theta_1) \geq 0$, for all $\mathbf{q}$. An agent's CPT utility in (4.5) can be written as:

$$\begin{aligned}\mathrm{CPT}(\mathbf{q}) =& W^+(p_n)[V(R(\mathbf{q},\theta_n)) - V(R(\mathbf{q},\theta_{n-1}))]+ \\ & W^+(p_{n-1}+p_n)[V(R(\mathbf{q},\theta_{n-1})) - V(R(\mathbf{q},\theta_{n-2}))] + \cdots + \\ & W^+\left(\sum_{x=2}^{n} p_x\right)[V(R(\mathbf{q},\theta_2)) - V(R(\mathbf{q},\theta_1))] + V(R(\mathbf{q},\theta_1))\end{aligned}$$

By keeping $\mathbf{q}$ fixed and differentiating the above function with respect to a probability value p_k, for some $k \in \{2,\ldots,n\}$, we obtain:

$$\begin{aligned}&\frac{\partial W^+\left(\sum_{x=k}^{n} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_k)) - V(R(\mathbf{q},\theta_{k-1}))]+ \\ &\frac{\partial W^+\left(\sum_{x=k-1}^{n} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_{k-1})) - V(R(\mathbf{q},\theta_{k-2}))] + \cdots + \\ &\frac{\partial W^+\left(\sum_{x=2}^{n} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_2)) - V(R(\mathbf{q},\theta_1))]\end{aligned}$$

Since W^+ is strictly increasing, $\frac{\partial W^+(\cdot)}{\partial p_k}$ is greater than zero. Moreover, since $R(\mathbf{q},\theta_n) > \cdots > R(\mathbf{q},\theta_1) \geq 0$ and V is strictly increasing, then $V(R(\mathbf{q},\theta_x)) - V(R(\mathbf{q},\theta_{x-1})) > 0$, for any $x \in \{2,\ldots,n\}$. Hence, the first derivative shown above is greater than zero, which implies that the CPT utility is strictly increasing in p_k when the other probabilities as well as $\mathbf{q}$ are fixed. Consequently, there are no stationary points and, thus, we map every reported belief to a single true belief. The right side of Figure 4.4 illustrates this.

If, alternatively, the scoring rule R allows $R(\mathbf{q},\theta_n) = R(\mathbf{q},\theta_{n-1}) = \cdots = R(\mathbf{q},\theta_1)$, then all partial derivatives illustrated above are equal to zero, which implies that there might exist multiple stationary points and, consequently, multiple true beliefs associated with a reported belief $\mathbf{q}$. Thus, we have shown how comonotonicity provides a sufficient condition under which positive proper scoring rules are indeed proper under CPT.

4.6.2 Negative Scoring Rules

Now, consider a negative proper scoring rule R that satisfies comonotonicity, *i.e.*, $0 \geq R(\mathbf{q},\theta_n) > \cdots > R(\mathbf{q},\theta_1)$, for all $\mathbf{q}$. An agent's CPT utility in (4.6) can then be written as:

$$\begin{aligned}\mathrm{CPT}(\mathbf{q}) =& W^-(p_1)[V(R(\mathbf{q},\theta_1)) - V(R(\mathbf{q},\theta_2))]+ \\ & W^-(p_1+p_2)[V(R(\mathbf{q},\theta_2)) - V(R(\mathbf{q},\theta_3))] + \cdots + \\ & W^-\left(\sum_{x=1}^{n-1} p_x\right)[V(R(\mathbf{q},\theta_{n-1})) - V(R(\mathbf{q},\theta_n))] + V(R(\mathbf{q},\theta_n))\end{aligned}$$

By keeping $\mathbf{q}$ fixed and differentiating the above function with respect to a probability value p_k, for some $k \in \{1,\ldots,n-1\}$, we obtain:

$$\begin{aligned}&\frac{\partial W^-\left(\sum_{x=1}^{k} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_k)) - V(R(\mathbf{q},\theta_{k+1}))]+ \\ &\frac{\partial W^-\left(\sum_{x=1}^{k+1} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_{k+1})) - V(R(\mathbf{q},\theta_{k+2}))] + \cdots + \\ &\frac{\partial W^-\left(\sum_{x=1}^{n-1} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_{n-1})) - V(R(\mathbf{q},\theta_n))]\end{aligned}$$

Since W^- is strictly increasing, $\frac{\partial W^-(\cdot)}{\partial p_k}$ is greater than zero. Moreover, since $0 \geq R(\mathbf{q},\theta_n) > R(\mathbf{q},\theta_{n-1}) > \cdots > R(\mathbf{q},\theta_1)$ and V is strictly increasing, then $V(R(\mathbf{q},\theta_x)) - V(R(\mathbf{q},\theta_{x+1})) > 0$, for any $x \in \{1,\ldots,n-1\}$. Hence, the first derivative shown above is greater than zero, which implies that the CPT utility is strictly increasing in p_k when the other probabilities as well as $\mathbf{q}$ are fixed. Consequently, we map every reported belief to a single true belief.

4.6.3 Mixed Scoring Rules

Finally, consider a mixed proper scoring rule R that satisfies comonotonicity, *i.e.*, $R(\mathbf{q},\theta_n) > \cdots > R(\mathbf{q},\theta_i) \geq 0 > R(\mathbf{q},\theta_{i-1}) > \cdots > R(\mathbf{q},\theta_1)$. An agent's CPT utility in (4.7) can then be written as:

$$
\begin{aligned}
\mathrm{CPT}(\mathbf{q}) =& W^+(p_n)[V(R(\mathbf{q},\theta_n)) - V(R(\mathbf{q},\theta_{n-1}))] + \cdots + \\
& W^+\left(\sum_{x=i+1}^{n} p_x\right)[V(R(\mathbf{q},\theta_{i+1})) - V(R(\mathbf{q},\theta_i))] + \\
& W^+\left(\sum_{x=i}^{n} p_x\right)V(R(\mathbf{q},\theta_i)) + W^-\left(\sum_{x=1}^{i-1} p_x\right)V(R(\mathbf{q},\theta_{i-1})) + \\
& W^-\left(\sum_{x=1}^{i-2} p_x\right)[V(R(\mathbf{q},\theta_{i-2})) - V(R(\mathbf{q},\theta_{i-1}))] + \cdots + \\
& W^-(p_1)[V(R(\mathbf{q},\theta_1)) - V(R(\mathbf{q},\theta_2))]
\end{aligned}
$$

If we keep $\mathbf{q}$ fixed and differentiate the above function with respect to a probability value p_k, for some $k \in \{i,\ldots,n\}$, we obtain:

$$
\begin{aligned}
& \frac{\partial W^+\left(\sum_{x=k}^{n} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_k)) - V(R(\mathbf{q},\theta_{k-1}))] + \cdots + \\
& \frac{\partial W^+\left(\sum_{x=i+1}^{n} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_{i+1})) - V(R(\mathbf{q},\theta_i))] + \\
& \frac{\partial W^+\left(\sum_{x=i}^{n} p_x\right)}{\partial p_k}V(R(\mathbf{q},\theta_i))
\end{aligned}
$$

Since W^+ is strictly increasing, $\frac{\partial W^+(\cdot)}{\partial p_k}$ is greater than zero. Moreover, since $R(\mathbf{q},\theta_n) > \cdots > R(\mathbf{q},\theta_i) \geq 0$ and V is strictly increasing, then $V(R(\mathbf{q},\theta_x)) - V(R(\mathbf{q},\theta_{x-1})) > 0$, for any $x \in \{i,\ldots,n\}$. Hence, the first derivative shown above is greater than zero.

If, on the other hand, $k \in \{1,\ldots,i-1\}$, we obtain:

$$
\begin{aligned}
& \frac{\partial W^-\left(\sum_{x=1}^{i-1} p_x\right)}{\partial p_k}V(R(\mathbf{q},\theta_{i-1})) + \\
& \frac{\partial W^-\left(\sum_{x=1}^{i-2} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_{i-2})) - V(R(\mathbf{q},\theta_{i-1}))] + \cdots + \\
& \frac{\partial W^-\left(\sum_{x=1}^{k} p_x\right)}{\partial p_k}[V(R(\mathbf{q},\theta_k)) - V(R(\mathbf{q},\theta_{k+1}))]
\end{aligned}
$$

Since W^- is strictly increasing, $\frac{\partial W^-(\cdot)}{\partial p_k}$ is greater than zero. Moreover, since $0 > R(\mathbf{q}, \theta_{i-1}) > \cdots > R(\mathbf{q}, \theta_1)$ and V is strictly increasing, then $V(R(\mathbf{q}, \theta_x)) - V(R(\mathbf{q}, \theta_{x+1})) > 0$, for any $x \in \{1, \ldots, i-1\}$. Hence, the first derivative shown above is also greater than zero. Consequently, we map every reported belief to a single true belief.

4.6.4 Constructing a Comonotonic Proper Scoring Rule

We finally discuss how to construct a comonotonic proper scoring rule S from a bounded proper scoring rule R with range (a, b). If R is a positive proper scoring rule, *i.e.*, $a, b \in \Re^+$, we then define a scoring rule S as:

$$S(\mathbf{q}, \theta_x) = R(\mathbf{q}, \theta_x) + (x-1) \times (b-a)$$

for $x \in \{1, \ldots, n\}$. The above scoring function has different non-overlapping ranges for different outcomes. Clearly, we obtain $S(\mathbf{q}, \theta_n) > \cdots > S(\mathbf{q}, \theta_1) \geq 0$, and S is trivially proper.

If, on the other hand, R is a negative proper scoring rule with range (a, b), for $a, b \in \Re^-$, we then define S as:

$$S(\mathbf{q}, \theta_x) = R(\mathbf{q}, \theta_x) + (n-x) \times (a-b)$$

for $x \in \{1, \ldots, n\}$. Once more, S is proper and it has different non-overlapping ranges for different outcomes. Moreover, we obtain $0 \geq S(\mathbf{q}, \theta_n) > \cdots > S(\mathbf{q}, \theta_1)$.

Finally, if R is a mixed proper scoring rule with range (a, b), for $a \in \Re^-$ and $b \in \Re^+$, we then define S as follows:

$$S(\mathbf{q}, \theta_x) = R(\mathbf{q}, \theta_x) + (n-x) \times (a-b), \text{ for } x \in \{1, \ldots, i-1\}$$
$$S(\mathbf{q}, \theta_x) = R(\mathbf{q}, \theta_x) + (x-1) \times (b-a), \text{ for } x \in \{i, \ldots, n\}$$

S is proper and $S(\mathbf{q}, \theta_n) > \cdots > S(\mathbf{q}, \theta_i) > 0 > S(\mathbf{q}, \theta_{i-1}) > \cdots > S(\mathbf{q}, \theta_1)$.

To summarize, the range of the score an agent receives from a comonotonic proper scoring rule depends on the observed outcome, where different non-overlapping ranges are associated with different outcomes. An agent's reported belief then defines the agent's score inside the possible ranges.

4.7 A General Method for Obtaining True Beliefs When the Utility Function, the Loss-Aversion Parameter, and the Weighting Functions Are Unknown

Thus far, we have focused on how to calibrate *a posteriori* an agent's reported belief when the same was elicited using a proper scoring rule. To do so, we have assumed that the components that drive the agent's attitude towards uncertainty are known, *i.e.*, the agent's utility function, loss-aversion parameter, and weighting functions were previously elicited. There are many techniques to elicit such components, *e.g.*, based on the trade-off method [1, 3, 90], and regret-based techniques [36, 91]. Clearly, eliciting those components incurs an extra burden since agents must be pre-screened before reporting their beliefs.

In this section, we propose a general method based on non-deterministic payments to obtain an agent's true belief when those components are unknown. Consider two fixed payoffs $s_{\max}$ and $s_{\min}$, where $s_{\max} > s_{\min} \geq 0$. The specific values for $s_{\max}$ and $s_{\min}$ are irrelevant in our analysis as long as the previous inequalities hold true. The scores from a proper scoring rule will now determine how likely it is that an agent will receive each fixed payoff. In other words, instead of defining the payoffs of a prospect (see Section 4.4), the scores from a proper scoring rule and, consequently, a reported belief, have now the dual role of determining the probabilities of obtaining the fixed payoffs. In order for the above interpretation to be true, we consider a positive proper scoring rule R with range $[0, 1]$. An agent's payment is defined as follows:

$$\begin{aligned}
&\textbf{If } \theta_1 \textbf{ happens: } \text{the agent receives the prospect } [s_{\min}\colon 1 - R(\mathbf{q}, \theta_1), s_{\max}\colon R(\mathbf{q}, \theta_1)];\\
&\textbf{If } \theta_2 \textbf{ happens: } \text{the agent receives the prospect } [s_{\min}\colon 1 - R(\mathbf{q}, \theta_2), s_{\max}\colon R(\mathbf{q}, \theta_2)];\\
&\qquad\qquad\qquad\qquad\vdots\\
&\textbf{If } \theta_n \textbf{ happens: } \text{the agent receives the prospect } [s_{\min}\colon 1 - R(\mathbf{q}, \theta_n), s_{\max}\colon R(\mathbf{q}, \theta_n)].
\end{aligned} \tag{4.15}$$

The above payment structure is based on the payment structure proposed by Sandroni and Shmaya [71]. However, Sandroni and Shmaya focused on the elicitation of beliefs under expected utility theory, whereas we focus on the elicitation of beliefs under cumulative prospect theory. The payment structure in (4.15) is essentially a compound prospect (or a two-stage prospect), *i.e.*, an agent receives the prospect $[s_{\min}\colon 1 - R(\mathbf{q}, \theta_1), s_{\max}\colon R(\mathbf{q}, \theta_1)]$ with probability p_1, the prospect $[s_{\min}\colon 1 - R(\mathbf{q}, \theta_2), s_{\max}\colon R(\mathbf{q}, \theta_2)]$ with probability p_2, and so on. We assume the reduction of compound lotteries axiom [70], *i.e.*, an agent finds

a two-stage prospect as attractive as the one-stage prospect yielding the same payoffs with the corresponding multiplied probabilities. In other words, the above compound prospect is equivalent to the following one-stage prospect:

$$\left[s_{\min}\colon \sum_{k=1}^{n} p_k(1 - R(\mathbf{q}, \theta_k)), s_{\max}\colon \sum_{k=1}^{n} p_k R(\mathbf{q}, \theta_k)\right] \tag{4.16}$$

An important feature of cumulative prospect theory is that it satisfies *stochastic dominance*, *i.e.*, a prospect $[s_{\min}\colon 1-\rho, s_{\max}\colon \rho]$ is preferred to a prospect $[s_{\min}\colon 1-\rho', s_{\max}\colon \rho']$ if and only if $\rho > \rho'$. Given that stochastic dominance holds true, we do not need to know how an agent evaluates the prospect in (4.16). Stochastic dominance implies that the agent will move as much probability mass towards the highest payoff $s_{\max}$ as possible. In other words, an agent reports $\mathbf{q}$ so as to maximize $\sum_{k=1}^{n} p_k R(\mathbf{q}, \theta_k)$, which is the probability of receiving the maximum payoff $s_{\max}$. Note that $\sum_{k=1}^{n} p_k R(\mathbf{q}, \theta_k)$ is the expected score from the proper scoring rule R, which implies that $\arg\max_{\mathbf{q}} \sum_{k=1}^{n} p_k R(\mathbf{q}, \theta_k) = \mathbf{p}$. That is, by using the payment structure in (4.15), we obtain an agent's true belief without having to elicit the components that drive the agent's attitude towards uncertainty.

4.8 Summary

Proper scoring rules have been widely used to induce honest reporting of subjective probabilities. The main assumption behind proper scoring rules is that agents are risk neutral, which is often an unrealistic assumption when the underlying agents are human beings. Our first contribution in this chapter was an experiment that corroborates the fact that human beings are not risk neutral even when the stakes are small. Different than other related experiments (*e.g.*, [92]), our experiment was performed in the context of elicitation of beliefs using proper scoring rules.

Thereafter, we adapted proper scoring rules to cumulative prospect theory [86], a more appropriate model of decision making under uncertainty. Given that CPT is descriptive in nature, it is then more empirically justifiable than normative theories such as expected-value maximization and expected-utility theory. In particular, we showed how to obtain an agent's true belief from his misreported belief by solving a system of equations. We covered all possible scenarios regarding the underlying proper scoring rule, *i.e.*, when the proper scoring rule is positive, negative, and mixed. Moreover, our discussion is valid for any finite number of outcomes, thus generalizing the work by Offerman *et al.* [61].

We also discussed how multiple true beliefs might be associated with a single reported belief when agents behave according to cumulative prospect theory. This implies that the reported belief cannot always be used to uniquely determine the correct true belief. The reason for this phenomenon is that traditional proper scoring rules do not satisfy comonotonicity. We provided a new explanation for why comonotonicity is a sufficient condition for fixing the above problem. We also suggested how to construct a comonotonic proper scoring rule from any traditional proper scoring rule, thus generalizing the work by Kothiyal *et al.* [45].

While the abovementioned contributions assumed that the components that drive an agent's attitude towards uncertainty are known *a priori*, we also proposed a new approach to elicit an agent's belief when those components are unknown. Our approach uses non-deterministic payments based on proper scoring rules. An important feature of our approach is that it is in the best interest of a CPT agent to behave honestly since this maximizes the agent's CPT utility.

4.8.1 Directions for Future Work

There are some interesting open questions regarding the applicability of our results. Offerman *et al.* [61] showed empirically that adapting the binary and positive version of the quadratic scoring rule to cumulative prospect theory results in a lower degree of violation of additivity in the reported probabilities. An interesting related question is whether or not adapting proper scoring rules to cumulative prospect theory results in more accurate reported beliefs, where accuracy is defined in terms of forecast accuracy. Another interesting question regards the effectiveness of our elicitation approach based on non-deterministic payments. In particular, it would be interesting to compare our proposed approach with traditional deterministic approaches in terms of how easy it is to apply and explain to the agents.

Finally, it would be of theoretical value to characterize how agents report their beliefs when only partial information about an agent's CPT utility is known. A potential approach to follow when only the value function is known is to incorporate the value function into the proper scoring rule, as suggested by Winkler and Murphy [95]. Formally, the composite function $S = V^{-1} \circ R$ is a proper scoring rule under V. In other words, the scoring rule $S(\mathbf{q}, \theta_x)$ is tailored to the agent's value function. It is easy to see from (4.5) and (4.6) that such tailored proper scoring rules elicit a vector of decision weights, which in turn is a probability vector for positive and negative scoring rules. However, after obtaining a vector of decision weights, one still needs to elicit the agent's weighting function in order

to remove the influence of the weighting function on the reported belief. Such a procedure can be performed similarly to the recursive approaches described in Sections 4.5.1, 4.5.2, and 4.5.3. We conjecture that no deterministic payment structure can elicit the true belief of a CPT agent when the agent's weighting functions are unknown.

Regarding the case when only the agent's weighting functions are known, we argue that this setting can be treated in a similar way to the case when one wants to elicit the true belief of an agent who behaves according to expected-utility theory. In particular, the probabilistic payments proposed in [4, 43, 71] would all elicit vectors of decision weights. Given that the agent's weighting functions are known, one can then perform simple recursive approaches to eliminate the influence of the weighting functions on the reported belief in a way similar to the procedures described in Sections 4.5.1, 4.5.2, and 4.5.3. We conjecture that no deterministic payment structure can elicit the true belief of a CPT agent when the agent's value functions are unknown.

The above conjectures find some support on the work by Lambert [46], who showed that without a non-deterministic payment structure, and without any knowledge of the agent's preferences, the only information that can be elicited with strict incentives is whether an event is more likely than its complement. All in all, we note that the non-deterministic payment structure proposed in Section 4.7 induces honest reporting when only partial information about the agent's CPT utility is available, as long as the reduction of compound lotteries axiom holds true.

Chapter 5

A Study on the Influence of the Number of Agents on the Quality of the Aggregate Information

While previous chapters focused mainly on how to induce honest reporting during the elicitation of private information, hereafter we focus on how to aggregate the reported information. When eliciting and aggregating agents' private information, arguments can be made in favor and against the use of multiple agents. On the one hand, multiple agents bring diversity to the process so that biases of individual judgments can offset each other, which may result in more accurate aggregate information. On the other hand, hiring a larger population of agents might be too costly.

In this chapter, we empirically investigate the above question through an experiment using one of the most popular crowdsourcing platforms: *Amazon's Mechanical Turk* (AMT). In particular, we use the data set resulting from our content-analysis experiment (see Section 3.5). Recall that in our experiment, we asked agents to solve three content-analysis tasks. Due to the nature of the tasks, we were able to derive *gold-standard answers*, *i.e.*, answers of high quality provided by experts with relevant expertise (see Appendix B).

The existence of gold-standard answers allows us to investigate how different combinations of agents affect the accuracy of aggregate answers. We first analyze the accuracy of aggregate answers as the number of agents increases. Focusing on simple averages to aggregate answers, we find a substantial degree of improvement in expected accuracy as we increase the number of agents, with diminishing returns for extra agents. Moreover, the standard deviation of errors in the aggregate answers decreases with more agents, which

implies less risk when aggregating answers.

Our experimental results also show that combining only the overall top-performing agents results in more accurate aggregate answers, and these agents are consistent across multiple homogeneous tasks. On the other hand, removing the overall worst-performing agents from the population of agents sometimes result in less accurate aggregate answers. The reason for this surprising result is that the overall worst-performing agents can produce good answers on some tasks, which implies that they tend to be inconsistent across multiple tasks. Our results thus contribute to a better understanding on how to design more effective crowdsourcing processes.

5.1 Related Work

Many different questions have been recently addressed by the crowdsourcing research community, *e.g.*, how to assign tasks to agents [37, 84], how to design optimal workflows to coordinate the work of the crowd [47, 98], how to induce honest behavior in crowdsourcing settings [31], *etc.*

To the best of our knowledge, our work in this chapter is the first one to address the question of how the number of agents affects the quality of aggregate answers in crowdsourcing settings. Similar studies have been performed in different domains. For example, it is well-known in decision analysis and operations research that combining multiple forecasts often leads to improved forecasting performance [23]. Sheng *et al.* [76] showed in a data mining/machine learning domain that labeling the same data set with different "labelers" might sometimes improve data quality.

However, some unexpected results are apparently specific to crowdsourcing. For example, our experimental results show that removing the overall worst-performing agents from the population of agents might result in less accurate aggregate answers. Thus, we expect our work to shed light on how to design more effective crowdsourcing processes.

5.2 The Content-Analysis Experiment

Recall that in our content-analysis experiment described in Section 3.5, each agent analyzed 3 texts, and reported 3 answers per text inside the set $\{0, 1, 2\}$, thus reporting a total of 9 answers. Since the domain of each answer is rather small, our analysis in this chapter is performed per text, *i.e.*, we consider triplets of answers. Henceforth, we denote by *output*

a vector of 3 answers for a given text. Thus, each agent reported 3 outputs. We define gold-standard outputs in a similar fashion. In our analysis, we use outputs from members of the group CG, *i.e.*, the control group that involved no incentives beyond the baseline compensation offered for completing the content-analysis task. We focus only on this group because that is how tasks are traditionally performed in crowdsourcing settings.

5.3 Accuracy of Aggregate Outputs by the Number of Agents

In this section, we study the influence of the number of agents on the quality of the aggregate output. In order to do so, we generated combinations of the 50 agents in our population. For $r \in \{1, \ldots, 4\}$ and $r \in \{46, \ldots, 50\}$, we calculated all possible combinations of agents, *i.e.*, $\binom{50}{r}$. For example, for $r = 2$, we generated all $\binom{50}{2} = 1225$ pairs of agents. Due to the intractable number of combinations for $r \in \{5, \ldots, 45\}$, we randomly generated 10^5 different combinations of agents for any $r \in \{5, \ldots, 45\}$.

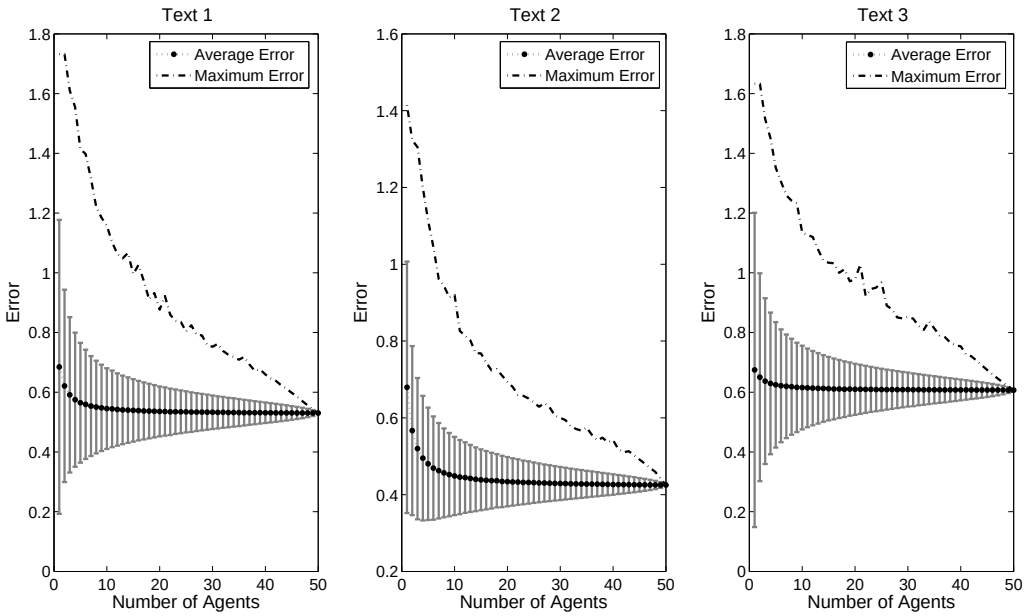
For each combination of agents, we aggregated the outputs from the underlying agents by taking the average of them. For instance, for two agents, we calculated the average output for all $\binom{50}{2} = 1225$ possible pairs of agents.

We then measured the accuracy of each aggregate output. For each aggregate output, we calculated the *root-mean-square deviation* (RMSD) between the aggregate output and the gold-standard output. For example, suppose that a pair of agents report the outputs $(1, 2, 0)$ and $(2, 2, 1)$ for Text 1. Thus, the aggregate output is $(1.5, 2, 0.5)$. Given that the gold-standard output for Text 1 is $(1, 2, 2)$ (see Appendix B), the root-mean-square deviation between the aggregate output and the gold-standard output is:

$$\sqrt{\frac{(1.5-1)^2 + (2-2)^2 + (0.5-2)^2}{3}} \approx 0.9129$$

We denote by *error* the RMSD between the aggregate output and the gold-standard output. Clearly, the lower the error, the more accurate the aggregate output. In our experiments, the range of the error is $[0, 2]$. The resulting *average error* for a given r can be seen as the *expected error* when aggregating outputs using r agents. For instance, the average of the $\binom{50}{2} = 1225$ errors from all possible pairs of agents is the *expected error* when aggregating outputs using 2 agents chosen at random. Figure 5.1 shows the average error, the standard deviation of the errors, and the maximum error per text for each $r \in \{1, \ldots, 50\}$. The complete numerical data is shown in Table C.1 in Appendix C.

Figure 5.1: The average error, the standard deviation of the errors, and the maximum error per text for each $r \in \{1, \ldots, 50\}$.

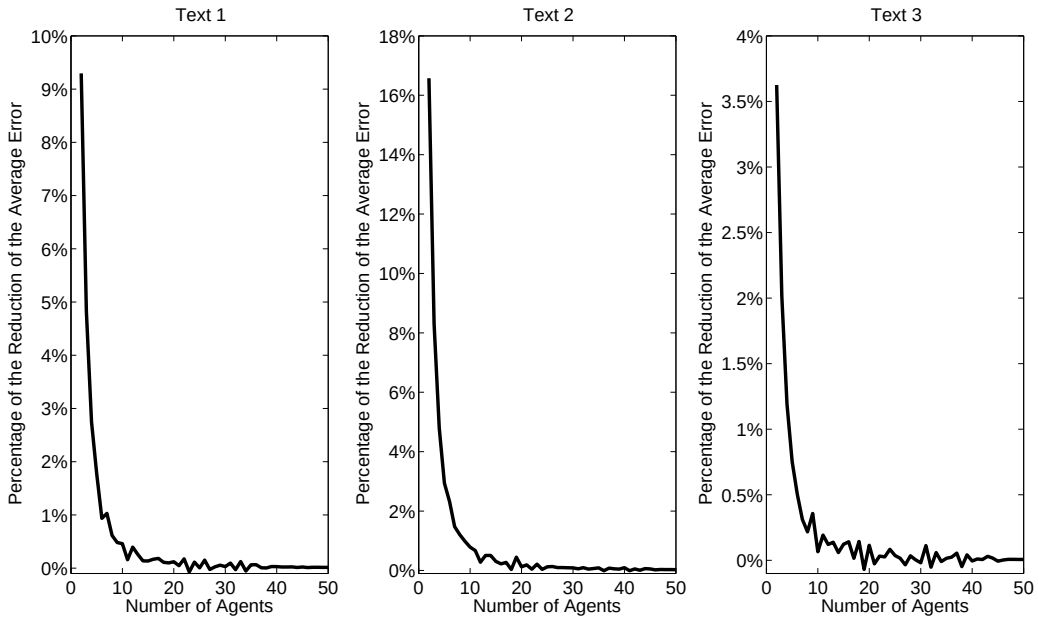


An interesting feature of Figure 5.1 is that the influence of the number of agents on the quality of the aggregate output is qualitatively the same for all texts. That is, the average error decreases as the number of agents r increases, which means that the expected accuracy of the aggregate output increases with more agents.

Figure 5.2 shows the percentage of the reduction of the average error when one extra agent is added. From the starting point of one agent, adding a second agent reduces the average error by $3.6\% - 16.5\%$. Given two agents, adding a third agent decreases the average error by $2\% - 8.3\%$, and so on. Clearly, there are diminishing returns for extra agents. For example, while adding a fourth agent reduces the average error by $1.19\% - 4.79\%$, adding a tenth agent reduces the average error by only $0.07\% - 0.79\%$. After the sixth agent, adding another agent always decreases the average error by less than 2% for all texts.

Figure 5.1 also shows that the standard deviation of the errors decreases with the number of agents r. The initially high standard deviation indicates an opportunity to get considerably low error with a single agent. Of course, the other side of the coin is a greater risk of high error with a single poor agent. As the number of agents increases, this risk

Figure 5.2: The percentage of the reduction of the average error when one extra agent is added to the population.



decreases because combinations of exclusively poor agents become less likely. This fact is also shown in the reduction of the maximum error when r increases, which implies less risk when aggregating outputs.

5.4 Accuracy of Outputs from the Top Agents

The analysis performed in the previous section is based on combinations of agents from the full population of agents. Two interesting follow-up questions are: 1) how much can accuracy be improved by restricting attention to combinations of the overall top-performing agents? and 2) how much can accuracy be improved by removing the overall worst-performing agents from the population of agents?

In order to answer these questions, we first sorted agents based on the *overall error*. Recall that each agent reported three outputs, each one consisting of three answers. We denote by *overall output* a vector of all nine reported answers. Likewise, we denote by

overall gold-standard output the vector of all nine answers from the gold-standard outputs. Then, the *overall error* of an agent is the RMSD between his overall output and the overall gold-standard output.

For example, suppose that an agent reports the following outputs for Text 1, 2, and 3: $(1, 2, 2)$, $(1, 2, 0)$, and $(1, 0, 0)$. Hence, his overall output is $(1, 2, 2, 1, 2, 0, 1, 0, 0)$. Recall that the gold-standard outputs for Text 1, 2, and 3 are, respectively, $(1, 2, 2)$, $(1, 2, 1)$, and $(0, 0, 0)$. Thus, the overall gold-standard output is $(1, 2, 2, 1, 2, 1, 0, 0, 0)$. Consequently, the agent's overall error is:

$$\sqrt{\frac{x}{9}} \approx 0.4714$$

where:

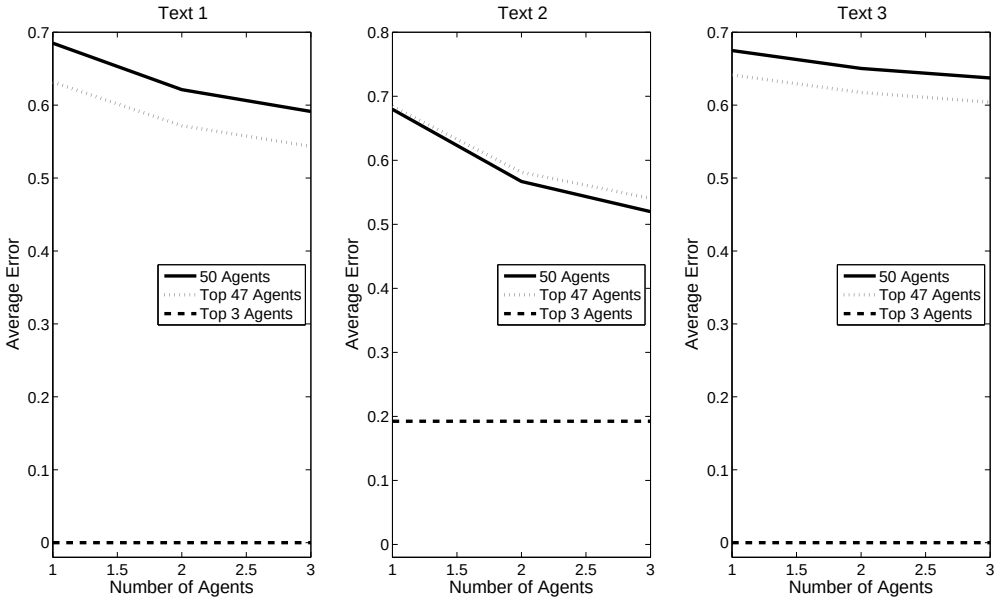
$$\begin{aligned} x =& (1-1)^2 + (2-2)^2 + (2-2)^2 + (1-1)^2 + (2-2)^2 + \\ & (0-1)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 \\ =& 2 \end{aligned}$$

For ease of exposition, in the following discussion we focus on the overall accuracy of the top 3 agents and on the accuracy of the population of agents without the 3 overall worst-performing agents, *i.e.*, the top 47 agents. We note, however, that the following results are qualitatively the same for the top k and the top $50-k$ agents, for any $k \in \{2, \ldots, 20\}$. We return to this point later in this section, when we also suggest a different way of ordering agents.

After ordering agents in terms of overall errors, we considered all possible combinations of the top 3 agents, *i.e.*, we calculated the aggregate outputs and errors for all $\binom{3}{r}$ possible combinations of agents, for $r \in \{1, 2, 3\}$. Moreover, we removed the three overall worst-performing agents from the full population of agents and calculated the aggregate outputs and errors for all $\binom{47}{r}$ combinations of agents, for $r \in \{1, 2, 3\}$ in order to allow quantitative comparisons across different populations of agents. The resulting average error per text for different populations of agents is illustrated in Figure 5.3. The complete numerical data is shown in Table C.1 in Appendix C.

Focusing first on Text 1 and 3, any combination of the top 3 agents results in a perfect aggregate output with zero error, whereas removing the three overall worst-performing

Figure 5.3: The average error per text for different populations of agents and $r \in \{1, 2, 3\}$.



agents reduces the average error by $4.96\% - 8.10\%$ in comparison with the complete population of agents, for the same group size $r \in \{1, 2, 3\}$.

Looking at the numerical values for Text 1 in Table C.1 (see Appendix C), the average error for combinations of 1, 2, and 3 agents from the top 47 agents (*i.e.*, 0.632, 0.572, and 0.543) is less than the average error for combinations of 1, 4, and 11 agents from the complete population of agents (*i.e.*, 0.685, 0.575, and 0.544). In other words, the aggregate outputs of 1, 2, and 3 randomly selected agents from the top 47 agents are expected to be more accurate in Text 1 than the aggregate outputs of 1, 4, and 11 randomly selected agents from the complete population of agents. These numbers for Text 3 are, respectively, 2, 8, and 50. Thus, for Text 1 and 3, it is beneficial to remove some worst-performing agents from the full population of agents.

The striking result comes from Text 2, where the average error for the full population of agents is $0.69\% - 3.85\%$ *lower* than the average error for the top 47 agents. The reason for this counter-intuitive result is that there were agents amongst the three overall worst-performing agents who excelled in Text 2, while performing poorly in Text 1 and 3. This shows that some agents are not consistent across multiple tasks. We return to this point in the next section.

For all populations of agents, the average error, the standard deviation of the errors, and the maximum error decrease as the number of agents increases, showing that combining multiple agents is always beneficial since it improves accuracy and reduces risks.

As mentioned before, for ease of exposition, our discussion in this section has been focused on the implications of restricting the population of agents to the overall top 3 agents and of removing the three overall worst-performing agents from the full population of agents. The obtained results are, however, more general. Any combination of up to k agents, for $k \in \{1, \ldots, 20\}$, from the top k agents results in a lower average error than a combination of the same number of agents from both the complete population of agents and the top $50-k$ agents. Moreover, removing any number $k \in \{2, \ldots, 20\}$ of worst-performing agents from the complete population of agents results in an increase of the average error for Text 2.

The above results are statistically significant for any $k \in \{3, \ldots, 20\}$ (rank-sum test, p-value ≤ 0.05). For combinations of size $k \in \{1, 2\}$, the three populations of agents have many combinations of agents in common. In general, as k increases, the fraction of combinations of agents shared between the top k agents, the top $50-k$ agents, and the full population of agents decreases, thus allowing us to make stronger statistical comparisons. For example, for $k \geq 4$, the p-values from the rank-sum tests are approximately 0.

It could be argued that the results in this section hold true because our experimental setting is biased, *e.g.*, the overall top-performing agents are expected to be more accurate in all texts because the overall error contains information about errors from all individual texts. However, if such a bias existed, combinations of top-performing agents would always result in lower average errors than combinations of the same number of agents from the full population of agents, a fact which is not true for $k \in \{21, \ldots, 25\}$. For example, for $k \in \{23, 24, 25\}$ and Text 1 and 2, a random combination of agents from the complete population of agents results in a lower average error than a random combination of the same number of agents from both the top k agents and the top $50 - k$ agents. In general, we find no clear pattern for values of $k \in \{21, \ldots, 25\}$.

Another way to compute the overall error and, thus, of ranking agents is by using a leave-one-out cross-validation approach. That is, given n texts, each agent receives a *historical rank* based on his errors on $n - 1$ texts. Then, the performance of different populations of agents is measured on the left-out text. However, the leave-one-out cross-validation approach may not work well with small data sets, such as the one in this study. We tried this approach on our data set and had mixed results. For example, when defining agents' historical ranks based on their performance in Text 1 and 2, and measuring the performance of different populations of agents in Text 3, a random combination of agents

from the top k agents resulted in *higher* average error than a random combination of the same number of agents from both the full population of agents and the top $50 - k$ agents, for some values of k. We conjecture that the above result is an artifact of having a small number of texts since the effect of a single text on the historical rank would likely be diluted if there was a larger number of texts.

To summarize, our results in this section imply that combining outputs from any number of the overall top 40% agents yields substantial improvements in expected accuracy in comparison to a combination of the same number of agents from the full population of agents, whereas removing agents amongst the overall worst 40% agents might result in less accurate aggregate outputs.

5.5 Consistency of Agents Across Multiple Tasks

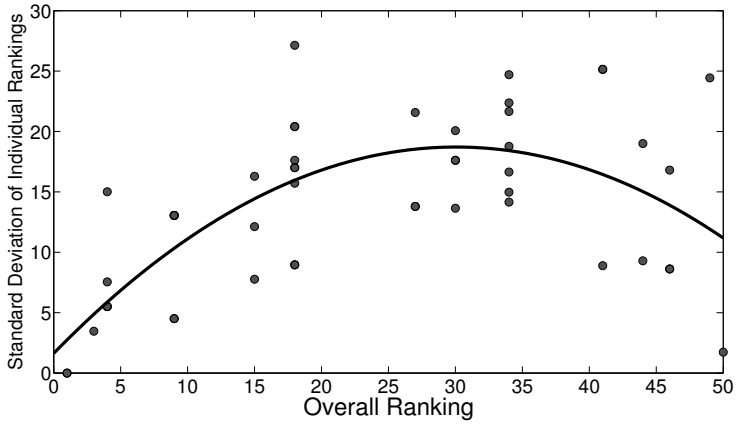
Our previous analysis shows that the relative performance of some agents is not necessarily consistent across multiple tasks. In order to further investigate this issue, we first calculated the *overall ranking* of agents in terms of overall errors, *i.e.*, we sorted agents in ascending order according to their overall errors.

Next, we calculated the *individual rankings* of each agent in terms of individual errors, *i.e.*, for each reported output, we sorted agents in ascending order according to their errors. Thus, each agent was ranked three times according to his errors. Ties in rankings were allowed, *i.e.*, agents with similar (overall) errors received the same ranking.

In the following analysis, we use the standard deviation of an agent's individual rankings as a measure of how stable the overall ranking of that agent is, where a high standard deviation indicates more ranking inconsistency across multiple tasks. For example, suppose that the outputs of an agent result in the lowest error in Text 1, the third lowest error in Text 2, and the second lowest error in Text 3. Thus, the standard deviation of that agent's individual rankings is equal to 1, showing high consistency across multiple tasks. On the other hand, an agent with individual rankings equal to 5, 48, and 22 is much more inconsistent across multiple tasks since the standard deviation of his individual rankings is equal to 21.66.

Figure 5.4 shows the standard deviation of individual rankings as a function of the overall ranking of each agent. For the sake of a better visualization, we fit a quadratic function to the data in a least-squares sense (norm of residuals equal to 35.664). We note that 2 is the optimal degree for polynomial fitting according to the Akaike information criterion (AIC). The resulting quadratic function is:

Figure 5.4: The standard deviation of individual rankings as a function of the overall ranking of each agent.



$$f(x) = -0.018922 * x^2 + 1.1371 * x + 1.6287$$

where x is an agent's overall ranking, and $f(x)$ is the standard deviation of that agent's individual rankings. Figure 5.4 shows that the overall top-performing agents are more consistent across multiple tasks than the other agents. For example, the standard deviations of the individual rankings of the top 7 agents are always less than 15, whereas 4 out of the 7 worst-performing agents have standard deviations greater than 15. In general, the most inconsistent agents are the agents with overall ranking between 15 and 35.

The results presented in this section, together with the results from the previous section, suggest that removing agents with high overall error from the population of agents might be a mistake since those agents can sometimes produce high quality outputs, as can be inferred from Figure 5.4. Furthermore, restricting the population of agents to a few overall top-performing agents is likely to produce more accurate aggregate outputs because these agents consistently report outputs with low errors.

5.6 Summary

In this chapter, we empirically studied the influence of the number of agents on the accuracy of aggregate outputs in a crowdsourcing setting. We first showed that having more agents reduces the average error of the aggregate output, which was measured in terms of the

root-mean-square deviation between the aggregate output and a gold-standard output. In other words, the expected accuracy of the aggregate output increases as the number of agents increases.

We also showed that there are diminishing returns for extra agents, where the reduction in the average error is always less than 2% after the sixth agent. Adding extra agents also implies that the risk of obtaining a combination of exclusively poor agents decreases because both the standard deviation of errors in aggregate outputs and the maximum error decrease as the number of agents increases.

We then moved to analyze the benefits of removing the overall worst-performing agents from the population of agents as well as the benefits of restricting the population of agents to only the overall top-performing agents. We found that an aggregate output from any combination of up to k top-performing agents, for $k \in \{1, \ldots, 20\}$, is, in expectation, more accurate than an aggregate output from a random combination with the same number of agents from the complete population of agents.

Unexpectedly, removing any number $k \in \{2, \ldots, 20\}$ of worst-performing agents does not necessarily result in more accurate aggregate outputs. The reason for this unexpected result is that the worst-performing agents are not always consistent across multiple tasks, which implies that a poor agent can eventually produce an accurate output.

Based on our results, our first recommendation for an organization or a decision maker who wants to design a crowdsourcing process is: in the absence of prior knowledge about the accuracy of the agents, having more agents is always beneficial because both the expected error in the aggregate output and the risk of obtaining a poor combination of agents decrease as the number of agents increases.

Clearly, the marginal costs as well as the marginal benefits of adding extra agents must be considered in practice. Our results showed that most of the benefit occurs with the first five to six agents. Thereafter, the marginal benefit of adding another agent is very low, and it might not outweigh the cost of adding the extra agent.

Our second recommendation for a more efficient design of crowdsourcing processes concerns the case when there exists prior knowledge about the accuracy of the agents. In this case, one should focus only on combinations of the overall top-performing agents since this greatly reduces the expected error in the aggregate output.

We found that almost perfect accuracy can be achieved by using only combinations of the very top agents. In practice, however, agents have constraints on the number of outputs they are willing to provide. This issue can be addressed by increasing the pool of top-performing agents. Our results show that when the size of the pool is up to 40% of

the size of the full population, the aggregate outputs from the top-performing agents are, in expectation, still more accurate than the aggregate outputs from the full population of agents, for the same number of agents.

5.6.1 Directions for Future Work

It is noteworthy that our study focused on simple averages to combine agents' outputs. Such an approach has been shown to perform well empirically and to be robust when eliciting expert opinions in different domains [23]. In addition, an averaging approach is easy to use, requiring neither assessments regarding the agent's judgment process nor self-assessed confidence in the accuracy of the reported outputs. However, it would be interesting to investigate whether or not we obtain similar results when using more sophisticated combination procedures [18].

Another exciting open question is whether or not the results obtained in our study hold true in different settings, *e.g.*, for different tasks, number of answers, *etc.* We argue that an answer to this question is of great importance to the crowdsourcing community given its potential to create less costly and more effective crowdsourcing processes. It would also be interesting to perform similar experiments, but under more contingent policies, *e.g.*, when the underlying workers are assessed online for quality and selected, rejected, or filtered to specific tasks as appropriate.

Finally, it would be of theoretical value to make stronger connections between our results in this section and standard results in statistics. For example, the metric mean square deviation can be stated in terms of variance and bias. That said, a promising approach is to analyze at which number of agents the bias stops decreasing. Such a number would be the ideal number of agents to hire since the statistical bias represents the true value of the parameter being estimated, *i.e.*, the aggregate error in our setting. Furthermore, the central limit theorem might help to derive stronger results. In particular, for each group size n, we obtain an x-bar distribution by considering the mean of each group instance as a data point. The analysis of the variance of such distributions might help to determine an optimal number of agents to hire, but now taking the risk of poor combinations of agents into account.

Chapter 6

A Consensual Linear Opinion Pool

Agents' private information usually take the form of either numerical point estimates or probability distributions over plausible outcomes. In this chapter, we focus on the latter case, *i.e.*, when agents report their beliefs once. We propose a pooling method to aggregate the reported beliefs into a single aggregate belief that works as if the agents were continuously updating their beliefs in order to accommodate the expertise and knowledge of others, *i.e.*, the proposed method mimics a consensus-reaching method. Each updated belief takes the form of a linear opinion pool, or a convex combination of beliefs, where the weight that an agent assigns to a peer's belief is inversely related to the distance between their beliefs. In other words, we assume agents prefer beliefs that are close to their own beliefs. We prove that such an updating process leads to consensus, *i.e.*, the agents all converge towards the same belief with probability one in the limit of the number of iterations, no matter their reported beliefs. We also show that if the beliefs of rational agents are scored using the quadratic scoring rule, then the assumption that agents prefer beliefs that are close to their own follows naturally.

6.1 Related Work

The aggregation of agents' private information have been extensively studied in computer science and, in particular, artificial intelligence, *e.g.*, the aggregation of private information represented as preferences over a set of alternatives as in social choice theory [22], the aggregation of point estimates using non-standard opinion pools [42], and the aggregation of subjective probabilities (beliefs) using prediction markets [21].

A traditional way of aggregating beliefs is through *opinion pooling methods*. These methods are often divided into behavioral and mathematical methods [24]. Behavioral aggregation methods attempt to generate agreement among the agents through interactions in order for them to share and exchange knowledge. Ideally, such sharing of information leads to a consensus. However, these methods typically provide no conditions under which the agents can be expected to reach agreement or even for terminating the iterative process.

On the other hand, mathematical aggregation methods consist of processes or analytical models that operate on the individual probability distributions in order to produce a single, aggregate probability distribution. An important mathematical method is the *linear opinion pool*, which involves taking a weighted linear average of the beliefs [25].

Several interpretations have been offered for the weights in the linear opinion pool. The performance-based approach recommends setting the weights based on previous performance of the agents [32]. A caveat with this approach is that performance measurements typically depend on the true outcome of the underlying event, which might not be available at the time when the beliefs have to be aggregated. Also, previous successful (respectively, unsuccessful) beliefs are not necessarily good indicators of future successful (respectively, unsuccessful) ones.

More closely related to our work is the interpretation of weights as a measure of distance. For example, Barlow *et al.* [6] proposed that the weight assigned to each agent's belief should be inversely proportional to its distance to the most distant belief, where distance is measured according to the Kullback-Leibler divergence. A clear drawback with this approach is that it only considers the distance to the most distant belief when assigning a weight to an agent's belief. Thus, even if the majority of agents have similar and accurate beliefs, the weights of these agents' beliefs in the aggregate belief can be greatly reduced due to a single distant belief.

For a comprehensive review of different perspectives on the weights in the linear opinion pool, we refer the interested reader to the work by Genest and McConway [32].

Even though we tackle the task of aggregating beliefs from a decision analysis perspective, we note that there is a field in machine learning called *ensemble learning* that deals with a similar problem, *i.e.*, how to use multiple learning algorithms to obtain better predictive performance than the performance achieved by any of the individual algorithms making up the ensemble [62]. We refer the reader interested in this perspective to the work by Dietterich [28].

6.2 The Model

We consider the setting where a decision maker is interested in a probability vector (belief) over a set of mutually exclusive outcomes $\theta_1, \ldots, \theta_n$, for $n \geq 2$. The decision maker deems it inappropriate to interject his own judgment about these outcomes. Hence, he elicits beliefs from z agents. Agents' beliefs are represented by n-dimensional probability vectors $\mathbf{q}_1, \ldots, \mathbf{q}_z$. The probability vector $\mathbf{q}_i = (q_{i,1}, \ldots, q_{i,n})$ represents agent i's belief, where $q_{i,k}$ is his subjective probability regarding the occurrence of outcome θ_k.

Since agents are not always in agreement, belief aggregation methods are used to combine their beliefs into a single probability vector. Formally, $\mathbf{q} = T(\mathbf{q}_1, \ldots, \mathbf{q}_z)$, where $\mathbf{q}$ is called an *opinion pool*, and the function T is the pooling operator. The *linear opinion pool* is a standard approach that involves taking a weighted linear average of the beliefs:

$$T(\mathbf{q}_1, \ldots, \mathbf{q}_z) = \sum_{i=1}^{z} w_i \mathbf{q}_i \tag{6.1}$$

where w_i denotes the weight associated with agent i's belief. We make the standard assumption that $0 \leq w_i \leq 1$, for every $i \in \{1, \ldots, z\}$, and $\sum_{i=1}^{z} w_i = 1$. Some axioms that support the use of the linear opinion pool have been proposed in the literature [53, 89].

6.2.1 Consensus and Weights

DeGroot [27] proposed a model which describes how a group can reach agreement on a common probability distribution by pooling their individual beliefs. Initially, each agent i is informed of the belief of every other agent. In order to accommodate the information and expertise of the rest of the group, agent i updates his own belief as follows:

$$\mathbf{q}_i^{(1)} = \sum_{j=1}^{z} w_{i,j} \mathbf{q}_j$$

where $w_{i,j}$ is the weight that agent i assigns to the belief of agent j when he carries out this update. Weights must be chosen on the basis of the relative importance that agents assign to their peers' beliefs. It is assumed that $w_{i,j} > 0$, for every agent i and j, and $\sum_{j=1}^{z} w_{i,j} = 1$. In this way, each updated belief takes the form of a linear opinion pool. The whole updating process can be written using matrix notation, *i.e.*, $\mathbf{Q}^{(1)} = \mathbf{W}\mathbf{Q}^{(0)}$, where:

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,z} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,z} \\ \vdots & \vdots & \ddots & \vdots \\ w_{z,1} & w_{z,2} & \cdots & w_{z,z} \end{bmatrix}, \quad \text{and} \quad \mathbf{Q}^{(0)} = \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \vdots \\ \mathbf{q}_z \end{bmatrix} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,n} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{z,1} & q_{z,2} & \cdots & q_{z,n} \end{bmatrix}$$

Since all the beliefs have changed, the agents might wish to revise their new beliefs in the same way as they did before. If there is no basis for the agents to change their weights, we can then represent the whole updating process after t revisions, for $t \geq 1$, as follows:

$$\mathbf{Q}^{(t)} = \mathbf{W}\mathbf{Q}^{(t-1)} = \mathbf{W}^t\mathbf{Q}^{(0)} \tag{6.2}$$

Let $\mathbf{q}_i^{(t)} = \left(q_{i,1}^{(t)}, \ldots, q_{i,n}^{(t)}\right)$ be agent i's belief after t updates, *i.e.*, it denotes the ith row of the matrix $\mathbf{Q}^{(t)}$. We say that a *consensus* is reached if $\mathbf{q}_i^{(t)} = \mathbf{q}_j^{(t)}$, for every agent i and j, as $t \to \infty$. Since $\mathbf{W}$, the matrix with weights, is a $z \times z$ stochastic matrix, it can then be regarded as the one-step transition probability matrix of a Markov chain with z states and stationary probabilities. Consequently, one can apply a limit theorem that says that a consensus is reached when there exists a positive integer t such that every element in at least one column of the matrix $\mathbf{W}^t$ is positive [27].

It is interesting to note that different forms of belief updating are possible without explicitly using other agents' disclosed beliefs. For example, in prediction markets, agents traditionally update their beliefs based on the current market price and market price movements (*e.g.*, see [49]).

6.2.2 Weights as a Measure of Distance

The original method proposed by DeGroot [27] has some drawbacks. First, the agents might want to change the weights that they assign to their peers' beliefs after learning their initial beliefs or after observing how much beliefs change from stage to stage. Further, beliefs and/or identities have to be disclosed to the whole group when agents are assigning weights. Hence, privacy is not preserved, a fact which might be troublesome when the underlying event is of a sensitive nature.

In order to tackle these problems, we derive weights that agents assign to the beliefs by interpreting each weight as a measure of distance between two beliefs. We start by

making the assumption that agents prefer beliefs that are close to their own beliefs, where closeness is measured by the following distance function:

$$D(\mathbf{q}_i, \mathbf{q}_j) = \sqrt{\frac{\sum_{k=1}^{n}(q_{i,k} - q_{j,k})^2}{n}} \tag{6.3}$$

i.e., it is the root-mean-square deviation between two beliefs $\mathbf{q}_i$ and $\mathbf{q}_j$. Given the above assumption, one can estimate the weight that agent i assigns to agent j's belief at a given time t, for $t \geq 1$, as follows:

$$w_{i,j}^{(t)} = \frac{\alpha_i^{(t)}}{\epsilon + D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_j^{(t-1)}\right)} \tag{6.4}$$

where $\alpha_i^{(t)}$ normalizes the weights so that they sum to one, and ϵ is a small, positive constant used to avoid division by zero. We set $\mathbf{q}_i^{(0)} = \mathbf{q}_i$, *i.e.*, it is the original belief reported by agent i. There are some important points regarding equation (6.4). First, the distance between two beliefs is always non-negative. Hence, the constant ϵ ensures that every single weight is strictly greater than 0 and strictly less than 1. Further, the closer the beliefs $\mathbf{q}_i^{(t-1)}$ and $\mathbf{q}_j^{(t-1)}$ are, the higher the resulting weight $w_{i,j}^{(t)}$ will be. Since $D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_i^{(t-1)}\right) = 0$, the weight that each agent assigns to his own belief is always greater than or equal to the weights that he assigns to his peers' beliefs. In spirit, the underlying learning model can be seen as a model of *anchoring* [85] in a sense that the belief of an agent is an "anchor", and subsequent updates are biased towards beliefs close to the anchor.

Now, we can redefine equation (6.2) so as to allow the agents to update their weights based on the most recent beliefs. After t revisions, for $t \geq 1$, we have that $\mathbf{Q}^{(t)} = \mathbf{W}^{(t)}\mathbf{Q}^{(t-1)} = \mathbf{W}^{(t)}\mathbf{W}^{(t-1)} \ldots \mathbf{W}^{(1)}\mathbf{Q}^{(0)}$, where each element of each matrix $\mathbf{W}^{(k)}$ is computed according to equation (6.4):

$$\mathbf{W}^{(k)} = \begin{bmatrix} w_{1,1}^{(k)} & w_{1,2}^{(k)} & \cdots & w_{1,z}^{(k)} \\ w_{2,1}^{(k)} & w_{2,2}^{(k)} & \cdots & w_{2,z}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{z,1}^{(k)} & w_{z,2}^{(k)} & \cdots & w_{z,z}^{(k)} \end{bmatrix}$$

The belief of each agent i at time t then becomes $\mathbf{q}_i^{(t)} = \sum_{j=1}^{z} w_{i,j}^{(t)} \mathbf{q}_j^{(t-1)}$. Algorithm 1 provides an algorithmic description of the proposed method.

Algorithm 1 Algorithmic description of the proposed method to find a consensual belief.

Require: z probability vectors $\mathbf{q}_1^{(0)}, \ldots, \mathbf{q}_z^{(0)}$.
Require: recalibration factor ϵ.
Require: number of rounds τ.
for $t = 1$ **to** τ **do**
  for $i = 1$ **to** z **do**
    for $j = 1$ **to** z **do**
      $w_{i,j}^{(t)} = \frac{\alpha_i^{(t)}}{\epsilon + D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_j^{(t-1)}\right)}$
    end for
    $\mathbf{q}_i^{(t)} = \sum_{j=1}^{z} w_{i,j}^{(t)} \mathbf{q}_j^{(t-1)}$
  end for
end for

In order to prove that all beliefs converge towards a consensual belief when using the proposed method, consider the following functions:

$$\delta\left(\mathbf{U}\right) = \frac{1}{2} \max_{i,j} \sum_{k=1}^{n} |u_{i,k} - u_{j,k}|$$

$$\gamma(\mathbf{U}) = \min_{i,j} \sum_{k=1}^{n} \min(u_{i,k}, u_{j,k})$$

where $0 \leq \delta\left(\mathbf{U}\right), \gamma(\mathbf{U}) \leq 1$, and $\mathbf{U}$ is a stochastic matrix. $\delta\left(\mathbf{U}\right)$ computes the maximum absolute difference between two rows of a stochastic matrix $\mathbf{U}$. Thus, when $\delta\left(\mathbf{Q}^{(t)}\right) = 0$, all rows of $\mathbf{Q}^{(t)}$ are the same, *i.e.*, a consensus is reached. We use the following results in our proof [63]:

Proposition 3. *Given two stochastic matrices* $\mathbf{U}$ *and* $\mathbf{V}$, $\delta(\mathbf{UV}) < \delta(\mathbf{U})\delta(\mathbf{V})$.

Proposition 4. *Given a stochastic matrix* $\mathbf{U}$, *then* $\delta(\mathbf{U}) = 1 - \gamma(\mathbf{U})$.

We state our main result below.

Proposition 5. *When* $t \to \infty$, $\mathbf{q}_i^{(t)} = \mathbf{q}_j^{(t)}$, *for every agent* i *and* j.

Proof. Recall that $\mathbf{Q}^{(t)}$ is the stochastic matrix representing the agents' beliefs after t revisions, and that $\mathbf{Q}^{(t)} = \mathbf{W}^{(t)}\mathbf{Q}^{(t-1)}$. Now, consider the following sequence:

$$\left(\delta\left(\mathbf{Q}^{(0)}\right), \delta\left(\mathbf{Q}^{(1)}\right), \ldots, \delta\left(\mathbf{Q}^{(t)}\right)\right)$$

We are interested in the behavior of this sequence when $t \to \infty$. First, we show that such a sequence is monotonically decreasing:

$$\begin{aligned}
\delta\left(\mathbf{Q}^{(t)}\right) &= \delta\left(\mathbf{W}^{(t)}\mathbf{Q}^{(t-1)}\right) \\
&< \delta\left(\mathbf{W}^{(t)}\right)\delta\left(\mathbf{Q}^{(t-1)}\right) \\
&= \left(1-\gamma\left(\mathbf{W}^{(t)}\right)\right)\delta\left(\mathbf{Q}^{(t-1)}\right) \\
&\leq \delta\left(\mathbf{Q}^{(t-1)}\right)
\end{aligned}$$

The second and third lines follow, respectively, from Propositions 3 and 4. Since $\delta(\mathbf{U}) \geq 0$, for every stochastic matrix $\mathbf{U}$, then the aforementioned sequence is a bounded decreasing sequence. Hence, we can apply the standard monotone convergence theorem [7] and $\delta\left(\mathbf{Q}^{(\infty)}\right) = 0$. Consequently, all rows of the stochastic matrix $\mathbf{Q}^{(\infty)}$ are the same.

To summarize, the sequence $\left(\delta\left(\mathbf{Q}^{(0)}\right), \delta\left(\mathbf{Q}^{(1)}\right), \ldots, \delta\left(\mathbf{Q}^{(t)}\right)\right)$ is strictly decreasing, which means that, according to the monotone convergence theorem, it will hit its tight lower bound after a potential infinite number of rounds. Given that the tight lower bound of $\delta(\cdot)$ is 0, then the sequence converges to 0. □

In words, a consensus is always reached under the proposed method, and this does not depend on the reported beliefs. A straightforward corollary of Proposition 5 is that all revised weights converge to the same value.

Corollary 1. *When $t \to \infty, w_{i,j}^{(t)} = \frac{1}{z}$, for every agent i and j.*

Hence, the proposed method works as if agents were continuously exchanging information so that their individual knowledge becomes group knowledge and all beliefs are equally weighted. Since we derive weights from the reported beliefs, we are then able to avoid some problems that might arise when eliciting these weights directly, *e.g.*, beliefs do not need to be disclosed to others in order for them to assign weights, thus preserving privacy.

The proposed method works very similarly to the behavioral aggregation method known as the Delphi method [48], where agents disclose and update their beliefs in a sequence of rounds in order to achieve consensus. In practice, however, consensus is not always

achieved under the Delphi method. By mimicking the Delphi method, our proposed algorithm always achieves consensus without requiring agents to explicitly participate in the consensus-reaching process. Hence, our proposed method is of great value whenever a requester is interested in a consensual belief, but achieving such a consensus might be costly in practice.

The resulting consensual belief can be represented as an instance of the linear opinion pool. Recall that $\mathbf{q}_i^{(t)} = \sum_{j=1}^{z} w_{i,j}^{(t)} \mathbf{q}_j^{(t-1)} = \sum_{j=1}^{z} w_{i,j}^{(t)} \sum_{k=1}^{z} w_{j,k}^{(t-1)} \mathbf{q}_k^{(t-2)} = \cdots = \sum_{j=1}^{z} \beta_j \mathbf{q}_j^{(0)}$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_z)$ is a probability vector that incorporates all the previous weights. Hence, another interpretation of the proposed method is that agents reach a consensus regarding the weights in equation (6.1).

6.2.3 Numerical Example

A numerical example may clarify the mechanics of the proposed method. Consider three agents ($z = 3$) with the following beliefs: $\mathbf{q}_1 = (0.9, 0.1)$, $\mathbf{q}_2 = (0.05, 0.95)$, and $\mathbf{q}_3 = (0.2, 0.8)$. According to (6.3), the initial distance between, say, $\mathbf{q}_1$ and $\mathbf{q}_2$ is:

$$D(\mathbf{q}_1, \mathbf{q}_2) = \sqrt{\frac{(0.9 - 0.05)^2 + (0.1 - 0.95)^2}{2}} = 0.85$$

Similarly, we have that $D(\mathbf{q}_1, \mathbf{q}_1) = 0$ and $D(\mathbf{q}_1, \mathbf{q}_3) = 0.7$. Using equation (6.4), we can then derive the weights that each agent assigns to the reported beliefs. Focusing on agent 1 at time $t = 1$ and setting $\epsilon = 0.01$, we obtain $w_{1,1}^{(1)} = \alpha_1^{(1)}/0.01$, $w_{1,2}^{(1)} = \alpha_1^{(1)}/0.86$, and $w_{1,3}^{(1)} = \alpha_1^{(1)}/0.71$. Since these weights must sum to one, we have $\alpha_1^{(1)} \approx 0.00975$ and, consequently, $w_{1,1}^{(1)} \approx 0.975$, $w_{1,2}^{(1)} \approx 0.011$, and $w_{1,3}^{(1)} \approx 0.014$. Repeating the same procedure for all agents, we obtain the matrix:

$$\mathbf{W}^{(1)} = \begin{bmatrix} 0.975 & 0.011 & 0.014 \\ 0.011 & 0.931 & 0.058 \\ 0.013 & 0.058 & 0.929 \end{bmatrix}$$

The updated belief of agent 1 is then $\mathbf{q}_1^{(1)} = \sum_{j=1}^{3} w_{1,j}^{(1)} \mathbf{q}_j \approx (0.8809, 0.1191)$. By repeating the above procedure, when $t \to \infty$, $\mathbf{W}^{(t)}$ converges to a matrix where all the elements are equal to $1/3$. Moreover, all agents' beliefs converge to the belief $(0.3175, 0.6825)$. An interesting point to note is that the resulting belief would be $(0.3833, 0.6167)$ if we had

taken the average of the reported beliefs, *i.e.*, agent 1, who has a very different belief, would have more influence on the aggregate belief.

6.3 Consensus and Proper Scoring Rules

The major assumption of the proposed method is that agents prefer beliefs that are close to their own beliefs. In this section, we formally investigate the validity of this assumption. We start by noting that in the absence of a well-chosen incentive structure, the agents might indulge in game playing which distorts their reported beliefs. For example, agents who have a reputation to protect might tend to produce beliefs near the most likely group consensus, whereas agents who have a reputation to build might tend to overstate the probabilities of outcomes they feel will be understated in a possible consensus [30]. As mentioned in Section 2.1, proper scoring rules are traditional devices used to promote honest reporting of beliefs

In particular, we consider effective scoring rules. Consider a metric G that assigns to any pair of beliefs $\mathbf{q}_i$ and $\mathbf{q}_j$ a real number, which in turn can be seen as the shortest distance between $\mathbf{q}_i$ and $\mathbf{q}_j$. Recall from Section 2.1 that a scoring rule R is *effective* with respect to G if the following relation holds for any beliefs $\mathbf{q}_i, \mathbf{q}_j$, and $\mathbf{q}_k$ [30]:

$$G(\mathbf{q}_i, \mathbf{q}_j) < G(\mathbf{q}_i, \mathbf{q}_k) \iff \mathbb{E}_{\mathbf{q}_i}\left[R(\mathbf{q}_j, \cdot)\right] > \mathbb{E}_{\mathbf{q}_i}\left[R(\mathbf{q}_k, \cdot)\right]$$

By definition, a metric G must satisfy the following conditions for any beliefs $\mathbf{q}_i, \mathbf{q}_j$, and $\mathbf{q}_k$:

1. Positivity: $G(\mathbf{q}_i, \mathbf{q}_j) \geq 0$, for all agents i, j, and $G(\mathbf{q}_i, \mathbf{q}_j) = 0$ if and only if $\mathbf{q}_i = \mathbf{q}_j$;
2. Symmetry: $G(\mathbf{q}_i, \mathbf{q}_j) = G(\mathbf{q}_j, \mathbf{q}_i)$;
3. Triangle Inequality: $G(\mathbf{q}_i, \mathbf{q}_k) \leq G(\mathbf{q}_i, \mathbf{q}_j) + G(\mathbf{q}_j, \mathbf{q}_k)$.

The root-mean-square deviation shown in (6.3) satisfies the above conditions. However, equation (6.4), taken as a function of beliefs, is not a true metric, *e.g.*, symmetry does not always hold. We adjust the original definition of effective scoring rules so as to consider weights instead of metrics. We say that a scoring rule R is effective with respect to a set of weights $\{w_{1,1}^{(t)}, \ldots, w_{1,z}^{(t)}, w_{2,1}^{(t)}, \ldots, w_{z,z}^{(t)}\}$, assigned at any time $t \geq 1$, if the following relation holds for any beliefs $\mathbf{q}_i^{(t-1)}, \mathbf{q}_j^{(t-1)}$, and $\mathbf{q}_k^{(t-1)}$:

$$w_{i,j}^{(t)} < w_{i,k}^{(t)} \iff \mathbb{E}_{\mathbf{q}_i^{(t-1)}}[R(\mathbf{q}_k^{(t-1)}, \cdot)] > \mathbb{E}_{\mathbf{q}_i^{(t-1)}}[R(\mathbf{q}_j^{(t-1)}, \cdot)]$$

In words, each agent's expected score can be seen as a monotone increasing function of his assigned weights, *i.e.*, the higher the weight one agent assigns to a peer's belief, the greater the expected score of that agent would be if he reported his peer's belief, and vice versa. We prove below that the quadratic scoring rule shown in (2.2) is effective with respect to a set of weights assigned according to (6.4).

Proposition 6. *The quadratic scoring rule shown in (2.2) is effective with respect to a set of weights* $\{w_{1,1}^{(t)}, \ldots, w_{1,z}^{(t)}, w_{2,1}^{(t)}, \ldots, w_{z,z}^{(t)}\}$ *assigned at any time* $t \geq 1$ *according to equation (6.4).*

Proof. Given a belief $\mathbf{q}_j$, we note that the $\mathbf{q}_i$-expected value of the quadratic scoring rule in (2.2) can be written as:

$$\begin{aligned}
\mathbb{E}_{\mathbf{q}_i}\left[R(\mathbf{q}_j, \cdot)\right] &= \sum_{e=1}^{n} q_{i,e}\, R(\mathbf{q}_j, \theta_e) \\
&= \sum_{e=1}^{n} \left(2q_{j,e}q_{i,e} - q_{i,e} \sum_{x=1}^{n} q_{j,x}^2 \right) \\
&= \sum_{e=1}^{n} 2q_{j,e}q_{i,e} - \sum_{e=1}^{n} q_{i,e} \sum_{x=1}^{n} q_{j,x}^2 \\
&= \sum_{e=1}^{n} 2q_{j,e}q_{i,e} - \sum_{x=1}^{n} q_{j,x}^2
\end{aligned}$$

for a set of exhaustive and mutually exclusive outcomes $\theta_1, \ldots, \theta_n$. Now, consider the weights assigned by agent i to the beliefs of agents j and k at time $t \geq 1$ according to equation (6.4). We have that $w_{i,j}^{(t)} < w_{i,k}^{(t)}$ if and only if:

$$
\begin{aligned}
&\frac{\alpha_i^{(t)}}{\epsilon + D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_j^{(t-1)}\right)} < \frac{\alpha_i^{(t)}}{\epsilon + D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_k^{(t-1)}\right)} && \equiv \\
&D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_k^{(t-1)}\right) < D\left(\mathbf{q}_i^{(t-1)}, \mathbf{q}_j^{(t-1)}\right) && \equiv \\
&\sum_{x=1}^{n}\left(q_{i,x}^{(t-1)} - q_{k,x}^{(t-1)}\right)^2 < \sum_{x=1}^{n}\left(q_{i,x}^{(t-1)} - q_{j,x}^{(t-1)}\right)^2 && \equiv \\
&\sum_{x=1}^{n} 2q_{i,x}^{(t-1)} q_{k,x}^{(t-1)} - \sum_{y=1}^{n}\left(q_{k,y}^{(t-1)}\right)^2 > \sum_{x=1}^{n} 2q_{i,x}^{(t-1)} q_{j,x}^{(t-1)} - \sum_{y=1}^{n}\left(q_{j,y}^{(t-1)}\right)^2 && \equiv \\
&\mathbb{E}_{\mathbf{q}_i^{(t-1)}}\left[R\left(\mathbf{q}_k^{(t-1)}, \cdot\right)\right] > \mathbb{E}_{\mathbf{q}_i^{(t-1)}}\left[R\left(\mathbf{q}_j^{(t-1)}, \cdot\right)\right]
\end{aligned}
$$

□

Proposition 6 implies that there is a correspondence between weights, assigned according to (6.4), and expected scores from the quadratic scoring rule: the higher the weight one agent assigns to a peer's belief, the greater that agent's expected score would be if he reported his peer's belief, and vice versa. Hence, whenever agents are rational, *i.e.*, when they behave so as to maximize their expected scores, and their beliefs are rewarded using the quadratic scoring rule, then the major assumption of the proposed method for finding a consensual belief, namely that agents prefer beliefs that are close to their own beliefs, is formally valid. A straightforward corollary of Proposition 6 is that a positive affine transformation of the quadratic scoring rule is still effective with respect to a set of weights assigned according to (6.4).

Corollary 2. *A positive affine transformation of the quadratic scoring rule R in (2.2), i.e., $xR(\mathbf{q}_i, \theta_e) + y$, for $x > 0$ and $y \in \Re$, is effective with respect to a set of weights $\{w_{1,1}^{(t)}, \ldots, w_{1,z}^{(t)}, w_{2,1}^{(t)}, \ldots, w_{z,z}^{(t)}\}$ assigned at any time $t \geq 1$ according to equation (6.4).*

It is important to clarify the implications of the above result. Proposition 6 means that if an agent receives a payment in a certain round according to the quadratic scoring rule, then it follows naturally that the agent should assign weights according to equation (6.4). In practice, however, agents are paid only once for the beliefs reported during the first round, since the algorithm behaves on their behalf from the second round on.

6.4 Empirical Evaluation

In this section, we describe an experiment designed to test the effectiveness of the proposed method for finding a consensual belief. In the following subsections, we describe the dataset used in our experiments, the metrics used to compare different methods to aggregate beliefs, and the obtained results.

6.4.1 Dataset

Our dataset was composed by 267 games (256 regular-season games and 11 playoff games) from the National Football League (NFL) held between September 8th, 2005 and February 5th, 2006. We obtained the beliefs of 519 agents for the NFL games from the ProbabilityFootball[1] contest. The contest was free to enter. Each agent was asked to report his subjective probability that a team would win a game. The beliefs had to be reported by noon on the day of the game. Since the probability of a tie in NFL games is very low (less than 1%), agents did not report the probability of such an outcome. In particular, no ties occurred in our dataset.

Not all 519 registered agents reported their beliefs for every game. An agent who did not enter a belief for a game was removed from the opinion pool for that game. On average, each game attracted approximately 432 agents, the standard deviation being equal to 26.37. The minimum and maximum number of agents were, respectively, 243 and 462. Importantly, the contest rewarded the performance of agents via a positive affine transformation of the quadratic scoring rule, *i.e.*, $100 - 400 \times p_l^2$, where p_l was the probability that an agent assigned to the eventual losing team.

Recall that a positive affine transformation of a strictly proper scoring rule is still strictly proper (Proposition 1). The above scoring rule can be obtained by multiplying the quadratic scoring rule in (2.2) by 200 and subtracting the result by 100. The resulting proper scoring rule rewards bold beliefs more when they are right. Likewise, it penalizes bold beliefs more when they are wrong. For example, a prediction of 99% earns 99.96 points if the chosen team wins, and it loses 292.04 points if the chosen team loses. On the other hand, a prediction of 51% earns 3.96 points if it is correct, and it loses 4.04 points if it is wrong. A prediction of 50% neither gains nor loses any points. The agents with highest accumulated scores won prizes in the contest. The suggested strategy at the contest website was "*to make picks for each game that match, as closely as possible, the probabilities that each team will win*".

[1] Available at http://probabilityfootball.com/2005/

We argue that this dataset is very suitable for our purposes due to many reasons. First, the popularity of NFL games provides natural incentives for people to participate in the ProbabilityFootball contest. Furthermore, the intense media coverage and scrutiny of the strengths and weaknesses of the teams and individual players provide useful information for the general public. Hence, participants of the contest can be viewed as knowledgeable regarding to the forecasting goal. Finally, the fact that agents were rewarded via a positive affine transformation of the quadratic scoring rule fits perfectly into the theory developed in this chapter (see Corollary 2).

6.4.2 Metrics

We used two different metrics to assess the prediction power of different aggregation methods.

Overall Accuracy

We say that a team is the predicted favorite for winning a game when an aggregate prediction that this team will win the game is greater than 0.5. Overall accuracy is then the percentage of games that predicted favorites have indeed won. A polling method with higher overall accuracy is more accurate.

Absolute Error

Absolute error is the difference between a perfect prediction (1 for the winning team) and the actual prediction. Thus, it is just the probability assigned to the losing team (p_l). An aggregate belief with lower absolute error is more accurate.

6.4.3 Experimental Results

For each game in our dataset, we aggregated the reported beliefs using three different linear opinion pools: the method we propose in this chapter, henceforth referred to as the *consensual* method; the traditional *average* approach, where all the weights in (6.1) are equal to $1/z$; and the method proposed by Barlow *et al.* [6], henceforth referred to as the *BMS* method. These authors proposed that the weight assigned to agent i's belief should be $w_i = \frac{c}{I(\mathbf{q}_i, \mathbf{q}_i^*)}$, where c is a normalizing constant, $I(\mathbf{q}_i, \mathbf{q}_i^*)$ is the Kullback-Leibler

Table 6.1: The average absolute error of each method over the 267 games. Standard deviations are in parentheses.

Consensual	Average	BMS
0.4115 (0.1813)	0.4176 (0.1684)	0.4295 (0.1438)

divergence, and $\mathbf{q}_{i^*}$ achieves $\max\{I(\mathbf{q}_i, \mathbf{q}_j) : 1 \leq j \leq z\}$, *i.e.*, $\mathbf{q}_i^*$ is the most distant belief from agent i's belief. The BMS method produces indeterminate outputs whenever there are probability assessments equal to 0 or 1. Hence, we recalibrated the reported beliefs when using the BMS method by replacing 0 and 1 by, respectively, 0.01 and 0.99.

After computing the aggregate beliefs, we calculated the performance of each method according to the accuracy metrics previously described. Regarding the overall accuracy of each method, the consensual method achieves the best performance in this experiment with an overall accuracy of 69.29%. The BMS and average methods achieve an overall accuracy of, respectively, 68.54% and 67.42%. Table 6.1 shows the average absolute error of each method over the 267 games. The consensual method achieves the best performance with an average absolute error of 0.4115. We performed left-tailed Wilcoxon signed-rank tests in order to investigate the statistical relevance of these results. The resulting p-values are all extremely small ($< 10^{-4}$), showing that the results are indeed statistically significant. One reason for such a small p-value is the fact that we used many data points in our analysis.

Despite displaying a decent overall accuracy, the BMS method has the worst performance according to the absolute error metric. A clear drawback with this method is that it only considers the distance to the most distant belief when assigning a weight to a belief. Since our experiments involve hundreds of agents, it is reasonable to expect at least one of them to have a very different and wrong belief.

The high number of agents should give an advantage to the average method since biases of individual judgment can offset with each other when beliefs are diverse, thus making the aggregate belief more accurate. However, the average method achieves the worst overall accuracy, and it performs statistically worse than the consensual method when measured under the absolute error metric. We believe this result happens because the average method ends up overweighting extreme beliefs when equally weighting all beliefs. On the other hand, under the consensual method, agents put less weight on beliefs far from their own beliefs, which implies that this method is generally less influenced by extreme beliefs as illustrated in Section 6.2.3.

6.5 Summary

We proposed a pooling method to aggregate beliefs. Intuitively, the proposed method works as if the agents were continuously updating their beliefs, where each updated belief takes the form of a linear opinion pool, and the weight that each agent assigns to a peer's belief is inversely related to the distance between their beliefs. We proved that this updating process leads to a consensus.

A different interpretation of the proposed method is that agents reach a consensus regarding the weights of a linear opinion pool. We showed that if rational agents are rewarded using the quadratic scoring rule, then our major assumption, namely that agents prefer beliefs that are close to their own beliefs, follows naturally. To the best of our knowledge, this is the first work linking the theory of proper scoring rules to the seminal consensus theory proposed by DeGroot [27].

Using real-world data, we compared the performance of the proposed method with two other methods: the traditional average approach and another distance-based aggregation method proposed by Barlow *et al.* [6]. The results of our experiment show that the proposed method outperforms all the other methods when measured in terms of both overall accuracy and absolute error.

6.5.1 Directions for Future Work

Given our positive empirical results, it would be interesting to compare how the proposed consensus-based aggregation method performs in comparison to other opinion polls as well as prediction markets. Form a theoretical perspective, an interesting research direction is to provide more theoretical guarantees for the proposed method. For example, it would be interesting to pinpoint the extent to which an extreme belief affects the aggregate belief under the consensual method.

Recall that the resulting consensual belief can be represented as an instance of the linear opinion pool, *i.e.*, $\mathbf{q}_i^{(t)} = \sum_{j=1}^{z} w_{i,j}^{(t)} \mathbf{q}_j^{(t-1)} = \sum_{j=1}^{z} w_{i,j}^{(t)} \sum_{k=1}^{z} w_{j,k}^{(t-1)} \mathbf{q}_k^{(t-2)} = \cdots = \sum_{j=1}^{z} \beta_j \mathbf{q}_j^{(0)}$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_z)$ is a probability vector that incorporates all the previous weights. Another interesting research direction is to characterize β in terms of $\mathbf{q}_1, \ldots, \mathbf{q}_z$, *i.e.*, to characterize the resulting consensus as a function of the agents' initial beliefs.

Chapter 7

Concluding Remarks

The major theme in this thesis was to look at ways of making the elicitation and aggregation of agents' private information more effective. We accomplished this goal in many distinct ways. In Chapter 3, we showed that, in the presence of social projection, one can induce honest reporting in settings that involve multiple-choice questions by making pairwise comparisons between reported answers and rewarding agreements. Further, we empirically showed that providing rewards through pairwise comparisons results in more accurate answers than when agents have no direct incentives for reporting their honest answers.

Our first contribution is of great relevance when eliciting private answers to multiple-choice questions due to its theoretical guarantees and easiness to use and explain to the underlying agents. Our work opens the door to many interesting research questions, *e.g.*:

- How can one induce social projection and, consequently, incentivize honest reporting under the proposed scoring method?
- Which variables affect the strength of social projection?
- Is there a correlation between the strength of social projection and expertise?

Focusing on the elicitation of beliefs, we experimentally showed in Chapter 4 that human beings fail to be risk neutral when reporting their beliefs under a proper scoring rule. We then discussed how reporting a belief under a proper scoring rule is equivalent to making a choice under uncertainty. We proposed a way to adapt proper scoring rules to cumulative prospect theory, a modern theory of choice under uncertainty. We explained why a property called comonotonicity is a sufficient condition for proper scoring rules to be

indeed proper under cumulative prospect theory. Moreover, we showed how to construct a comonotonic proper scoring rule from any traditional proper scoring rule. Finally, we also proposed a new approach that uses non-deterministic payments based on proper scoring rules to elicit an agent's true belief.

Our contributions in Chapter 4 improve the elicitation process by considering a more modern and, arguably, accurate decision theory when using proper scoring rules. An exciting research direction is to investigate how much of improvement one obtain by assuming that agents follow cumulative prospect theory, rather than displaying risk-neutral behavior.

Moving to the aggregation of agents' reports, we empirically investigated in Chapter 5 how the number of workers on the crowdsourcing platform Amazon's Mechanical Turk influences the quality of the aggregate answers in a content-analysis task. We found that both the expected error in the aggregate answer as well as the risk of a poor combination of workers decrease as the number of workers increases. Moreover, our results showed that restricting the population of workers to up to the overall top 40% workers is likely to produce more accurate aggregate answers, whereas removing up to the overall worst 40% workers can actually make the aggregate answer less accurate. We found that this result holds due to top-performing workers being consistent across multiple tasks, whereas worst-performing workers tend to be inconsistent.

Our results in Chapter 5 have very relevant practical implications. For example, a requester might not need to hire dozens of workers to complete a task. Our results show that after the 6th worker, the marginal benefit of having another worker is very low. An interesting research direction regarding our study in Chapter 5 concerns investigating whether similar results hold true in different settings, *e.g.*, for different tasks.

Finally, in Chapter 6, we proposed a pooling method to aggregate agents' beliefs. Intuitively, the proposed method works as if the agents were continuously updating their beliefs to accommodate the expertise of others. Each updated belief takes the form of a linear opinion pool, where the weight that an agent assigns to a peer's belief is inversely related to the distance between their beliefs. We proved that such an updating process leads to consensus. Further, we showed that if risk-neutral agents are rewarded using the quadratic scoring rule, then the assumption that they prefer beliefs that are close to their own beliefs follows naturally. We empirically demonstrated the effectiveness of the proposed method using real-world data. Given that the algorithm proposed in Chapter 6 mimics a consensus-reaching process, an interesting research direction is to investigate how its results compare to empirical consensus-reaching processes such as the Delphi method [48].

References

[1] M. Abdellaoui. Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science*, 46(11):1497–1512, 2000.

[2] M. Abdellaoui, H. Bleichrodt, and O. L'Haridon. A Tractable Method to Measure Utility and Loss Aversion Under Prospect Theory. *Journal of Risk and Uncertainty*, 36(3):245–266, 2008.

[3] M. Abdellaoui, H. Bleichrodt, and C. Paraschiv. Loss Aversion Under Prospect Theory: A Parameter-Free Measurement. *Management Science*, 53(10):1659–1674, 2007.

[4] F. Allen. Discovering Personal Probabilities When Utility Functions are Unknown. *Management Science*, 33(4):542–544, 1987.

[5] D. F. Bacon, Y. Chen, I. Kash, D. C. Parkes, M. Rao, and M. Sridharan. Predicting Your Own Effort. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 695–702, 2012.

[6] R. E. Barlow, R. W. Mensing, and N. G. Smiriga. Combination of experts' opinions based on decision theory. In A. P. Basu, editor, *Reliability and quality control*, pages 9–19. North-Holland, 1986.

[7] R. G. Bartle and D. R. Sherbert. *Introduction to Real Analysis*. Wiley, 3rd edition, 2000.

[8] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, 1994.

[9] J. E. Bickel. Scoring Rules and Decision Analysis Education. *Decision Analysis*, 7(4):346–357, 2010.

[10] C. Boutilier. Eliciting forecasts from self-interested experts: scoring rules for decision makers. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 737–744, 2012.

[11] L. Brenner and B. Bilgin. Preference, Projection, and Packing: Support Theory Models of Judgments of Others Preferences. *Organizational Behavior and Human Decision Processes*, 115(1):121–132, 2011.

[12] M. D. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[13] J. R. Busemeyer and E. M. Pothos. Social Projection and a Quantum Approach for Behavior in Prisoner's Dilemma. *Psychological Inquiry*, 23(1):28–34, 2012.

[14] A. Carvalho, S. Dimitrov, and K. Larson. The Output-Agreement Method Induces Honest Behavior in the Presence of Social Projection. *ACM SIGecom Exchanges*, 13(1):77–81, 2014.

[15] A. Carvalho and K. Larson. Sharing a Reward Based on Peer Evaluations. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 1455–1456, 2010.

[16] A. Carvalho and K. Larson. A Truth Serum for Sharing Rewards. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pages 635–642, 2011.

[17] A. Carvalho and K. Larson. Sharing Rewards Among Strangers Based on Peer Evaluations. *Decision Analysis*, 9(3):253–273, 2012.

[18] A. Carvalho and K. Larson. A Consensual Linear Opinion Pool. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2518–2524, 2013.

[19] Y. Chen, X. A. Gao, R. Goldstein, and I. A. Kash. Market manipulation with outside incentives. *Autonomous Agents and Multi-Agent Systems*, pages 1–36, 2014.

[20] Y. Chen, I. Kash, M. Ruberry, and V. Shnayder. Decision markets with good incentives. In *Internet and Network Economics*, pages 72–83. Springer, 2011.

[21] Y. Chen and D.M. Pennock. Designing markets for prediction. *AI Magazine*, 31(4):42–52, 2010.

[22] Y. Chevaleyre, U. Endriss, J. Lang, and N. Maudet. A short introduction to computational social choice. In *Proceedings of the 33rd conference on Current Trends in Theory and Practice of Computer Science*, pages 51–69, 2007.

[23] R. T Clemen. Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989.

[24] R. T. Clemen and R. L. Winkler. Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis*, 19:187–203, 1999.

[25] R.M. Cooke. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, 1991.

[26] R. M. Dawes. Statistical Criteria for Establishing a Truly False Consensus Effect. *Journal of Experimental Social Psychology*, 25(1):1–17, 1989.

[27] M. H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[28] T. G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[29] E. S. Epstein. A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*, 8(6):985–987, 1969.

[30] D. Friedman. Effective Scoring Rules for Probabilistic Forecasts. *Management Science*, 29(4):447–454, 1983.

[31] X. A. Gao, A. Mao, and Y. Chen. Trick or Treat: Putting Peer Prediction to the Test. In *Proceedings of the 1st Workshop on Crowdsourcing and Online Behavioral Experiments*, 2013.

[32] C. Genest and K.J. McConway. Allocating the weights in the linear opinion pool. *Journal of Forecasting*, 9(1):53–73, 1990.

[33] I. Gilboa. Expected Utility with Purely Subjective Non-Additive Probabilities. *Journal of mathematical Economics*, 16(1):65–88, 1987.

[34] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[35] R. Hanson. Combinatorial Information Market Design. *Information Systems Frontiers*, 5(1):107–119, 2003.

[36] G. Hines and K. Larson. Preference elicitation for risky prospects. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 889–896, 2010.

[37] C.-J. Ho and J. W. Vaughan. Online Task Assignment in Crowdsourcing Markets. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 45–51, 2012.

[38] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics*, 14(3):399–425, 2011.

[39] S.-W. Huang and W.-T. Fu. Enhancing Reliability Using Peer Consistency Evaluation in Human Computation. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 639–648, 2013.

[40] P. G. Ipeirotis. Analyzing the Amazon Mechanical Turk Marketplace. *XRDS Crossroads: The ACM Magazine for Students*, 17(2):16–21, 2010.

[41] V. R. Jose. A Characterization for the Spherical Scoring Rule. *Theory and Decision*, 66(3):263–281, 2009.

[42] R. Jurca and B. Faltings. Incentives for expressing opinions in online polls. In *Proceeddings of the 2008 ACM Conference on Electronic Commerce*, pages 119–128, 2008.

[43] E. Karni. A Mechanism for Eliciting Probabilities. *Econometrica*, 77(2):603–606, 2009.

[44] D. Katz and F. H. Allport. *Students' attitudes*. 1931.

[45] A. Kothiyal, V. Spinu, and P. P. Wakker. Comonotonic Proper Scoring Rules to Measure Ambiguity and Subjective Beliefs. *Journal of Multi-Criteria Decision Analysis*, 17(3-4):101–113, 2011.

[46] N. Lambert. Probability elicitation for agents with arbitrary risk preferences. Working paper, Graduate School of Business, Stanford University, 2011.

[47] C. H. Lin and D. S. Weld. Dynamically Switching Between Synergistic Workflows for Crowdsourcing. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 132–133, 2012.

[48] H. A. Linstone and M. Turoff. *The Delphi method: Techniques and applications*. Addison-Wesley Reading, MA, 1975.

[49] C. F. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91(3):425–429, 2006.

[50] M. Marge, S. Banerjee, and A. I. Rudnicky. Using the Amazon Mechanical Turk for Transcription of Spoken Language. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing*, pages 5270–5273, 2010.

[51] G. Marks and N. Miller. Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review. *Psychological Bulletin*, 102(1):72, 1987.

[52] W. Mason and S. Suri. Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2012.

[53] K. J. Mcconway. Marginalization and linear opinion pools. *Journal of the American Statistical Association*, 76(374):410–414, 1981.

[54] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, 51(9):1359–1373, 2005.

[55] B. Mullen, J. L. Atkins, D. S. Champion, C. Edwards, D. Hardy, J. E. Story, and M. Vanderklok. The False Consensus Effect: A Meta-Analysis of 115 Hypothesis Tests. *Journal of Experimental Social Psychology*, 21(3):262–283, 1985.

[56] A. H. Murphy. A Note on the Ranked Probability Score. *Journal of Applied Meteorology*, 10(1):155–156, 1970.

[57] Y. Nakazono. Strategic Behavior of Federal Open Market Committee Board Members: Evidence from Members' Forecasts. *Journal of Economic Behavior & Organization*, 93:62–70, 2013.

[58] R. F. Nau. Should Scoring Rules Be "Effective"? *Management Science*, 31(5):527–535, 1985.

[59] A. J. Nederhof. Methods of Coping with Social Desirability Bias: A Review. *European Journal of Social Psychology*, 15(3):263–280, 1985.

[60] P. Neruda. *100 Love Sonnets*. Exile, Bilingual edition, 2007.

[61] T. Offerman, J. Sonnemans, G. Van De Kuilen, and P. P. Wakker. A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes. *The Review of Economic Studies*, 76(4):1461–1489, 2009.

[62] D. Opitz and R. Maclin. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[63] A. Paz. *Introduction to Probabilistic Automata*. Academic Press, 1971.

[64] S. Plous. *The Psychology of Judgment and Decision Making*. Mcgraw-Hill Book Company, 1993.

[65] D. Prelec. A Bayesian Truth Serum for Subjective Data. *Science*, 306(5695):462–466, 2004.

[66] J. Quiggin. A Theory of Anticipated Utility. *Journal of Economic Behavior & Organization*, 3(4):323–343, 1982.

[67] G. Radanovic and B. Faltings. A Robust Bayesian Truth Serum for Non-Binary Signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 833–839, 2013.

[68] J. M. Robbins and J. I. Krueger. Social Projection to Ingroups and Outgroups: A Review and Meta-Analysis. *Personality and Social Psychology Review*, 9(1):32–47, 2005.

[69] L.; Ross, D.; Green, and P. House. The "False in Social Consensus Perception Effect": An Egocentric Bias and Attribution Processes. *Journal of Experimental Social Psychology*, 13(3):279–301, 1977.

[70] P. A. Samuelson. Probability, utility, and the independence axiom. *Econometrica*, pages 670–678, 1952.

[71] A. Sandroni and E. Shmaya. Eliciting Beliefs by Paying in Chance. *Economic Theory Bulletin*, 1(1):33–37, 2013.

[72] L. J. Savage. Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[73] D. Schmeidler. Subjective Probability and Expected Utility without Additivity. *Econometrica*, 57(3):571–587, 1989.

[74] R. Selten. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics*, 1(1):43–62, 1998.

[75] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 275–284, 2011.

[76] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

[77] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[78] D. J. Spiegelhalter. Probabilistic Prediction in Patient Management and Clinical Trials. *Statistics in Medicine*, 5(5):421–433, 1986.

[79] C.-A. S. Staël von Holstein. A Family of Strictly Proper Scoring Rules Which Are Sensitive to Distance. *Journal of Applied Meteorology*, 9(3):360–364, 1970.

[80] C. Starmer. Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk. *Journal of Economic Literature*, 38(2):332–382, 2000.

[81] J. Surowiecki. *The Wisdom of Crowds*. Random House LLC, 2005.

[82] J. Taylor, A. Taylor, and K. Greenaway. *Little Ann and Other Poems*. Nabu Press, 2010.

[83] P. E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

[84] L. Tran-Thanh, S. Stein, A. Rogers, and N. R. Jennings. Efficient Crowdsourcing of Unknown Experts Using Multi-Armed Bandits. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 768–773, 2012.

[85] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.

[86] A. Tversky and D. Kahneman. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[87] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[88] B. Waggoner and Y. Chen. Information Elicitation Sans Verification. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content*, 2013.

[89] Carl Wagner. Allocation, lehrer models, and the consensus of probabilities. *Theory and Decision*, 14:207–220, 1982.

[90] P. Wakker and D. Deneffe. Eliciting von Neumann-Morgenstern Utilities When Probabilities are Distorted or Unknown. *Management Science*, 42(8):1131–1150, 1996.

[91] T. Wang and C. Boutilier. Incremental utility elicitation with the minimax regret decision criterion. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 309–318, 2003.

[92] B. J. Weber and G. B. Chapman. Playing for peanuts: Why is risk seeking more common for low-stakes gambles? *Organizational Behavior and Human Decision Processes*, 97(1):31–46, 2005.

[93] R. R. J. Weiss. *Optimally Aggregating Elicited Expertise: A Proposed Application of the Bayesian Truth Serum for Policy Analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[94] R. L. Winkler and A. H. Murphy. "Good" Probability Assessors. *Journal of Applied Meteorology*, 7(5):751–758, 1968.

[95] R. L. Winkler and A. H. Murphy. Nonlinear Utility and the Probability Score. *Journal of Applied Meteorology*, 9:143–148, 1970.

[96] J. Witkowski and D. C. Parkes. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

[97] Y. Yinon, A. Mayraz, and S. Fox. Age and the False-Consensus Effect. *The Journal of Social Psychology*, 134(6):717–725, 1994.

[98] H. Zhang, E. Horvitz, and D. Parkes. Automated Workflow Synthesis. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1020–1026, 2013.

APPENDICES

Appendix A

Dirichlet Distributions

For our purposes, the *Dirichlet distribution* is as a continuous distribution over parameter vectors of a categorical distribution. Since $\boldsymbol{\omega}$ is the unknown parameter of the categorical distribution that models the population knowledge, then it is natural to consider a Dirichlet distribution as a prior for $\boldsymbol{\omega}$. Given a *hyperparameter* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, which is a vector of positive reals that determines the shape of the Dirichlet distribution, the probability density function of the Dirichlet distribution over $\boldsymbol{\omega}$ is:

$$P(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{k=1}^{n} \omega_k^{\alpha_k - 1} \tag{A.1}$$

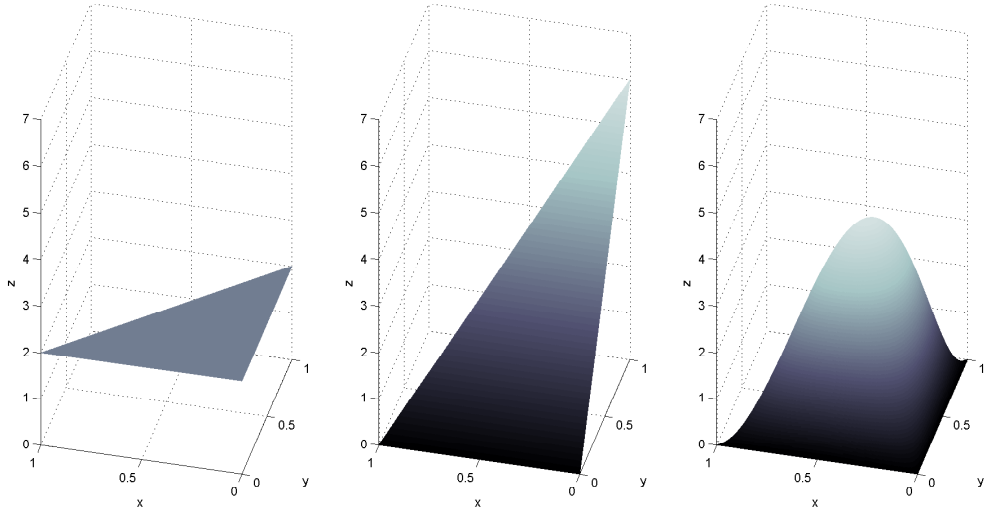
where:

$$\beta(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{n} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{n} \alpha_k\right)}$$

and Γ is the Gamma function. Figure A.1 shows the above probability density when $n = 3$ for some parameter vectors $\boldsymbol{\alpha}$. For the Dirichlet distribution in (A.1), the expected value of ω_j is $\mathbb{E}[\omega_j|\boldsymbol{\alpha}] = \alpha_j / \sum_{k=1}^{n} \alpha_k$. The probability vector $\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}] = (\mathbb{E}[\omega_1|\boldsymbol{\alpha}], \ldots, \mathbb{E}[\omega_n|\boldsymbol{\alpha}])$ is called the *expected distribution* regarding $\boldsymbol{\omega}$.

An interesting property of the Dirichlet distribution is that it is the *conjugate prior* of the categorical distribution [8], *i.e.*, the posterior distribution $P(\boldsymbol{\omega}|\boldsymbol{\alpha}, t_i)$ is itself a Dirichlet distribution. This relationship is often used in Bayesian statistics to estimate hidden parameters of categorical distributions. To illustrate this point, suppose that agent i

Figure A.1: Probability densities of Dirichlet distributions when $n = 3$ for different parameter vectors. Left: $\boldsymbol{\alpha} = (1, 1, 1)$. Center: $\boldsymbol{\alpha} = (2, 1, 1)$. Right: $\boldsymbol{\alpha} = (2, 2, 2)$.



observes the signal $t_i = A_x$, for $x \in \{1, \ldots, n\}$. After applying Bayes' rule, agent i's posterior distribution is $P(\boldsymbol{\omega}|\boldsymbol{\alpha}, t_i = A_x) = P(\boldsymbol{\omega}|(\alpha_1, \alpha_2, \ldots, \alpha_x+1, \ldots \alpha_n))$. Consequently, the new expected distribution regarding $\boldsymbol{\omega}$ is:

$$\mathbb{E}[\boldsymbol{\omega}|\boldsymbol{\alpha}, t_i = A_x] = \left(\frac{\alpha_1}{1 + \sum_{k=1}^{n} \alpha_k}, \ldots, \frac{\alpha_x + 1}{1 + \sum_{k=1}^{n} \alpha_k}, \ldots, \frac{\alpha_n}{1 + \sum_{k=1}^{n} \alpha_k} \right) \tag{A.2}$$

We call the probability vector in (A.2) agent i's *posterior predictive distribution* regarding $\boldsymbol{\omega}$ because it provides the distribution of future answers given the observed signal t_i. Consequently, the values $\alpha_1, \ldots, \alpha_n$ are "pseudo-counts" from "pseudo-data", where each α_k is the number of times that the ω_k-probability event has been observed before.

A Dirichlet distribution is *uninformative* when all the elements making up the vector $\boldsymbol{\alpha}$ have the same value. Uninformative priors are used when there is no prior knowledge favoring one probability event over another. The assumption of uninformative priors is reasonable in Chapter 3 because there is no ground for one agent to favor one answer over another *a priori*, *i.e.*, before observing and solving the multiple-choice question.

By using Dirichlet distributions as priors, belief updating can be expressed as an updating of the parameters of the prior distribution. Furthermore, the assumption of social projection allows the center to estimate agents' posterior distributions based solely on their reported answers, a point which is explored by our scoring method in Chapter 3.

Appendix B

Content-Analysis Experiment

In this appendix, we describe the texts used in our content-analysis experiment as well as the gold-standard answers reported by five professors and tutors from the English and Literature Department at the University of Waterloo, henceforth referred to as the *experts*.

Text 1

An excerpt from the "Sonnet XVII" by Neruda [60]. Intentionally misspelled words are highlighted in bold.

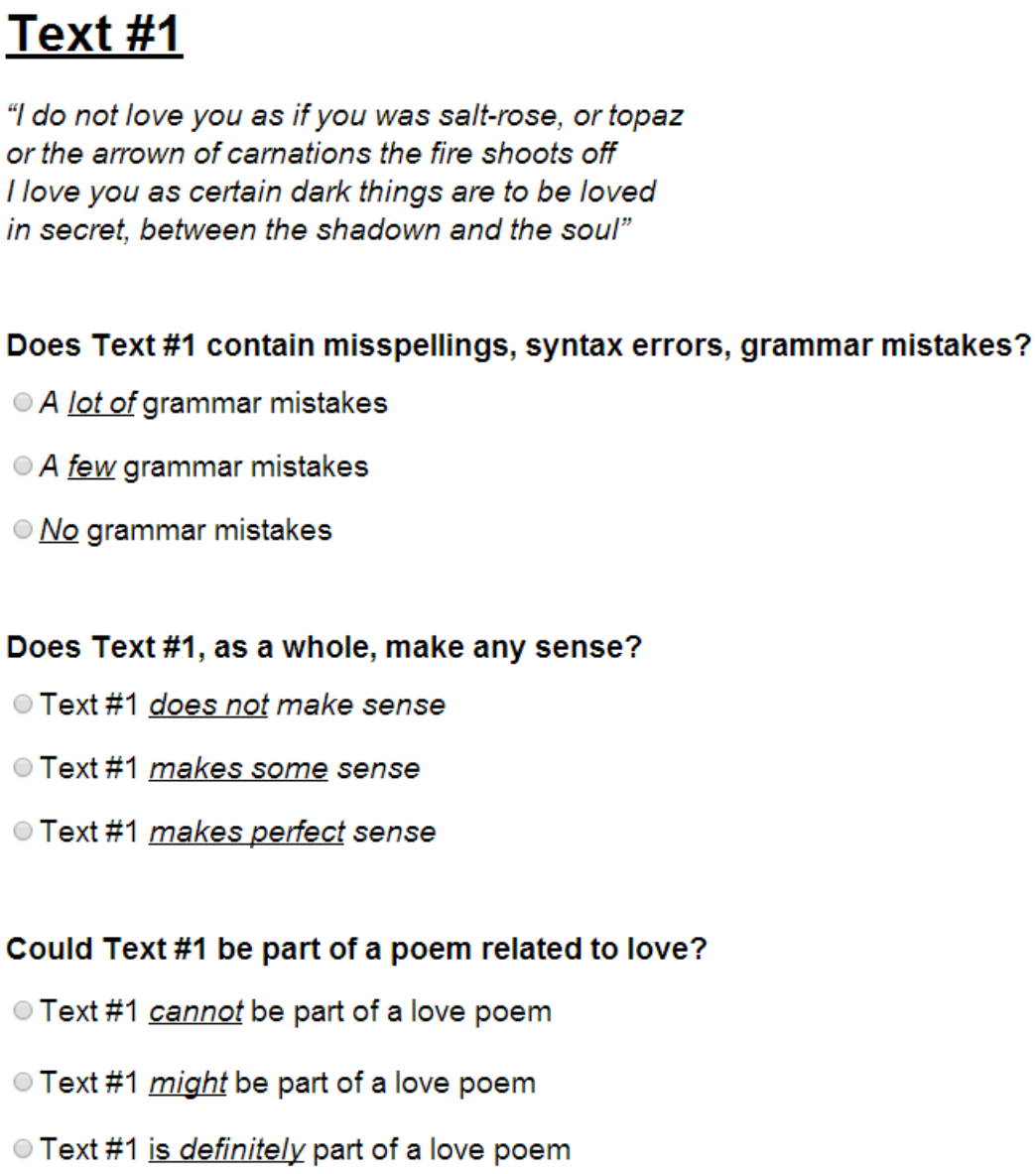
> "I do not love you as if you **was** salt-rose, or topaz,
> or the **arrown** of carnations that spread fire:
> I love you as certain dark things are loved,
> secretly, between the **shadown** and the soul"

Figure B.1 shows the graphical interface used in our experiment. Table B.1 shows the answers reported by the experts. The gold-standard answer for each criterion is the median/mode of the experts' answers.

Table B.1: Answers reported by the experts for Text 1.

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Median/Mode
Grammar	1	0	1	0	1	1
Clarity	2	2	2	1	2	2
Relevance	2	2	2	2	2	2

Figure B.1: Graphical interface used in our content-analysis experiment to elicit answers for Text 1.



Text 2

An excerpt from "The Cow" by Taylor *et al.* [82]. Intentionally misspelled words are highlighted in bold.

> "THANK you, **prety** cow, that made
> **Plesant** milk to soak my bread,
> Every day and every night,
> Warm, and fresh, and sweet, and white."

Figure B.2 shows the graphical interface used in our experiment. Table B.2 shows the answers reported by the experts. The gold-standard answer for each criterion is the median/mode of the experts' answers.

Table B.2: Answers reported by the experts for Text 2.

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Median/Mode
Grammar	1	1	1	1	1	1
Clarity	2	2	2	1	2	2
Relevance	1	0	0	1	1	1

Text 3

Random words in a semi-structured way. Each line starts with a noun followed by a verb in a wrong verb form. All the words in the same line start with a similar letter in order to mimic a poetic writing style.

> "Baby bet binary boundaries bubbles
> Carlos cease CIA conditionally curve
> Daniel deny disease domino dumb
> Faust fest fierce forced furbished"

Figure B.3 shows the graphical interface used in our experiment. Table B.3 shows the answers reported by the experts. The gold-standard answer for each criterion is the median/mode of the experts' answers.

Figure B.2: Graphical interface used in our content-analysis experiment to elicit answers for Text 2.

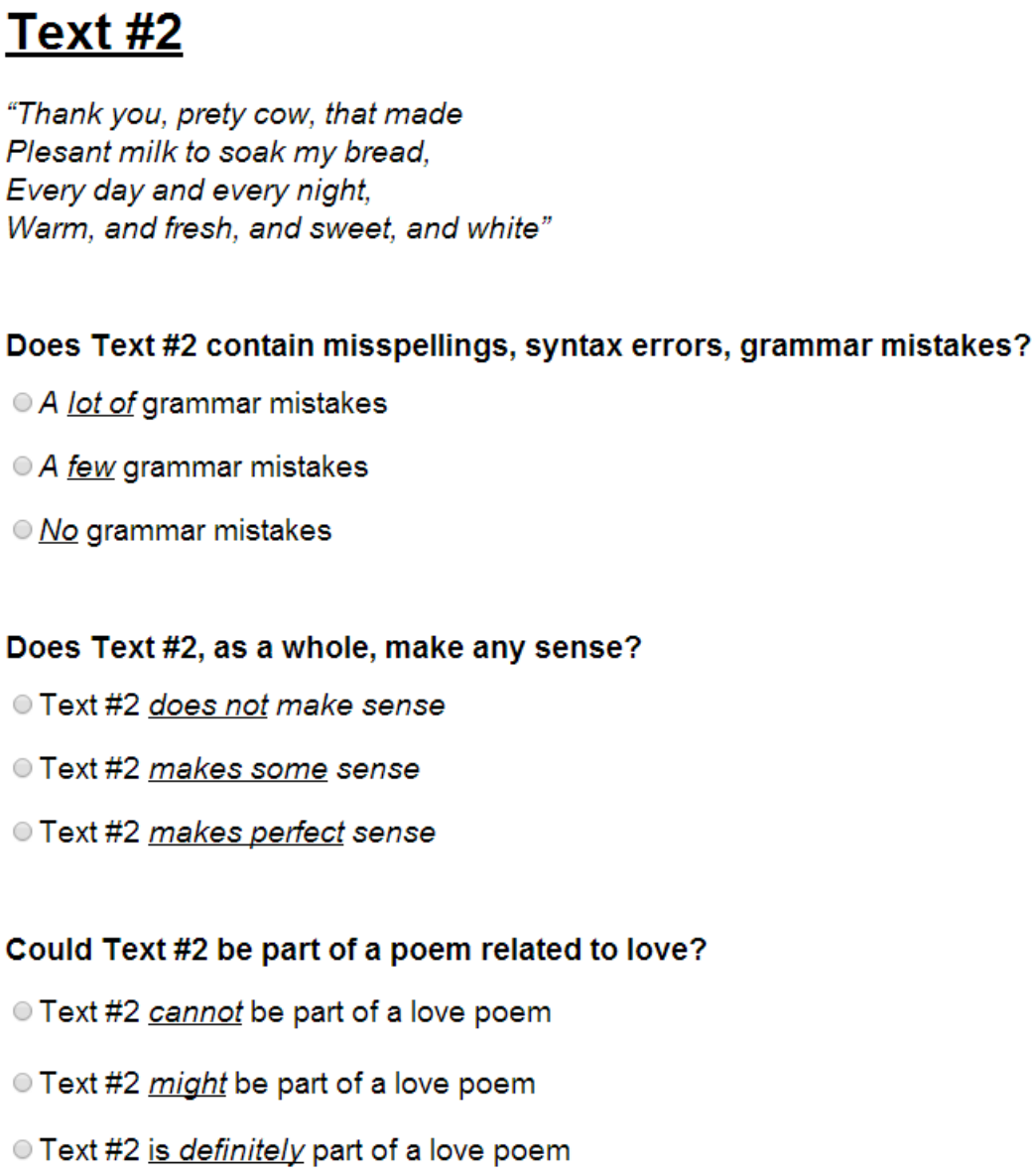


Figure B.3: Graphical interface used in our content-analysis experiment to elicit answers for Text 3.

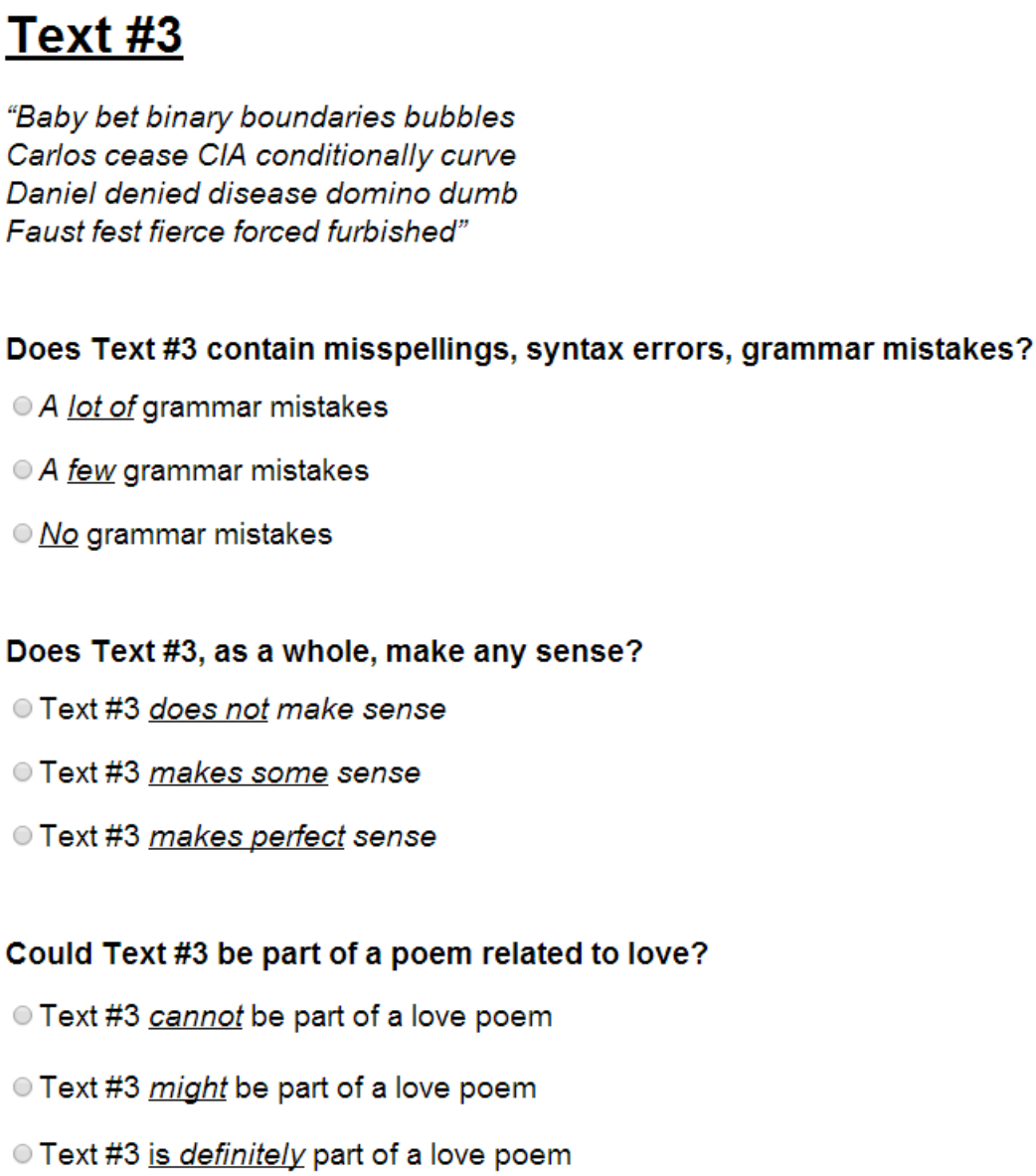


Table B.3: Answers reported by the experts for Text 3.

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Median/Mode
Grammar	0	1	0	0	0	0
Clarity	0	0	0	0	0	0
Relevance	0	1	0	0	0	0

Appendix C

Numerical Results of Chapter 5

Table C.1 shows the numerical results from all the analysis performed in Chapter 5.

Table C.1: The average error, the standard deviation of the errors, and the maximum error per text for different populations of agents. All the values are rounded to 3 decimal places.

		Text 1			Text 2			Text 3		
Population	r	Avg	Std	Max	Avg	Std	Max	Avg	Std	Max
All	1	0.685	0.492	1.732	0.679	0.327	1.414	0.675	0.526	1.633
	2	0.621	0.322	1.732	0.567	0.220	1.323	0.650	0.348	1.633
	3	0.591	0.260	1.610	0.520	0.184	1.305	0.637	0.278	1.515
	4	0.575	0.225	1.555	0.495	0.162	1.199	0.630	0.237	1.451
	5	0.565	0.201	1.414	0.480	0.146	1.114	0.625	0.210	1.352
	6	0.560	0.183	1.398	0.469	0.135	1.045	0.622	0.189	1.305
	7	0.554	0.167	1.314	0.462	0.124	0.962	0.620	0.173	1.259
	8	0.550	0.155	1.220	0.457	0.116	0.944	0.618	0.161	1.242
	9	0.548	0.145	1.185	0.452	0.109	0.914	0.616	0.150	1.232
	10	0.545	0.135	1.158	0.449	0.102	0.920	0.616	0.140	1.134
	11	0.544	0.128	1.102	0.446	0.096	0.827	0.615	0.132	1.127
	12	0.542	0.121	1.063	0.444	0.092	0.814	0.614	0.124	1.120
	13	0.541	0.115	1.048	0.442	0.087	0.802	0.613	0.118	1.078
	14	0.540	0.110	1.067	0.440	0.083	0.768	0.613	0.111	1.042
	15	0.539	0.105	0.996	0.439	0.079	0.768	0.612	0.106	1.034

Continued on the next page

		Text 1			Text 2			Text 3		
Population	r	Avg	Std	Max	Avg	Std	Max	Avg	Std	Max
	16	0.538	0.100	1.026	0.438	0.076	0.744	0.611	0.102	1.032
	17	0.537	0.096	0.973	0.436	0.073	0.724	0.611	0.097	1.000
	18	0.537	0.092	0.911	0.436	0.069	0.726	0.610	0.093	1.012
	19	0.536	0.088	0.932	0.434	0.067	0.710	0.611	0.089	0.971
	20	0.536	0.084	0.877	0.434	0.064	0.694	0.610	0.086	0.979
	21	0.535	0.081	0.922	0.433	0.062	0.681	0.610	0.082	1.027
	22	0.535	0.078	0.860	0.433	0.059	0.660	0.610	0.079	0.923
	23	0.535	0.074	0.839	0.432	0.057	0.657	0.610	0.075	0.946
	24	0.534	0.072	0.835	0.432	0.055	0.650	0.609	0.073	0.950
	25	0.534	0.069	0.804	0.431	0.053	0.640	0.609	0.070	0.967
	26	0.534	0.066	0.824	0.431	0.051	0.629	0.609	0.067	0.890
	27	0.534	0.064	0.793	0.430	0.048	0.638	0.609	0.065	0.878
	28	0.534	0.061	0.791	0.430	0.047	0.627	0.609	0.062	0.850
	29	0.533	0.059	0.758	0.429	0.045	0.606	0.609	0.059	0.847
	30	0.533	0.056	0.752	0.429	0.043	0.600	0.609	0.057	0.852
	31	0.533	0.054	0.762	0.429	0.041	0.594	0.608	0.055	0.847
	32	0.533	0.052	0.738	0.428	0.040	0.578	0.609	0.052	0.823
	33	0.532	0.049	0.726	0.428	0.038	0.571	0.608	0.050	0.809
All	34	0.532	0.047	0.718	0.428	0.036	0.567	0.608	0.048	0.837
	35	0.532	0.045	0.709	0.428	0.035	0.575	0.608	0.046	0.813
	36	0.532	0.043	0.718	0.428	0.033	0.556	0.608	0.044	0.791
	37	0.532	0.041	0.692	0.427	0.031	0.542	0.608	0.041	0.785
	38	0.532	0.039	0.672	0.427	0.030	0.548	0.608	0.039	0.767
	39	0.531	0.036	0.677	0.427	0.028	0.537	0.608	0.037	0.757
	40	0.531	0.034	0.660	0.426	0.027	0.542	0.608	0.035	0.754
	41	0.531	0.032	0.647	0.426	0.025	0.518	0.608	0.033	0.727
	42	0.531	0.030	0.638	0.426	0.023	0.510	0.608	0.030	0.720
	43	0.531	0.028	0.623	0.426	0.021	0.513	0.607	0.028	0.706
	44	0.531	0.026	0.612	0.426	0.020	0.502	0.607	0.026	0.690
	45	0.531	0.023	0.598	0.426	0.018	0.491	0.607	0.023	0.675
	46	0.531	0.020	0.585	0.426	0.016	0.481	0.607	0.021	0.660
	47	0.531	0.017	0.570	0.425	0.013	0.470	0.607	0.018	0.646
	48	0.531	0.014	0.556	0.425	0.011	0.459	0.607	0.014	0.632
	49	0.530	0.010	0.543	0.425	0.008	0.442	0.607	0.010	0.620

Continued on the next page

		Text 1			Text 2			Text 3		
Population	r	Avg	Std	Max	Avg	Std	Max	Avg	Std	Max
	50	0.530	0.000	0.530	0.425	0.000	0.425	0.607	0.000	0.607
Top 47	1	0.632	0.456	1.732	0.684	0.318	1.414	0.641	0.525	1.633
	2	0.572	0.298	1.555	0.581	0.219	1.323	0.617	0.346	1.633
	3	0.543	0.241	1.503	0.541	0.182	1.305	0.604	0.275	1.515
Top 3	1	0.000	0.000	0.000	0.192	0.333	0.577	0.000	0.000	0.000
	2	0.000	0.000	0.000	0.192	0.167	0.289	0.000	0.000	0.000
	3	0.000	0.000	0.000	0.192	0.000	0.192	0.000	0.000	0.000