UNIVERSITY OF SOUTHAMPTON

FACULTY OF ECONOMIC, SOCIAL AND POLITICAL SCIENCES

Department of Economics

Essays in Bounded Rationality and Economic Experiments

by

Lunzheng Li

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

in the

Faculty of Economic, Social and Political Sciences

Department of Economics

Jun 2020

University of Southampton

Abstract

Faculty of Economic, Social and Political Sciences

Economics

Thesis for the degree of Doctor of Philosophy

Essays in Bounded Rationality and Economic Experiments

by

Lunzheng Li

This thesis studies bounded rationality in three different contexts through the lens of economic experiments. The first two essays focus on preferences, which constitute the foundation of economic theory. In modern economics, preferences are assumed to have a consistent underlying structure based on a small number of axioms. However, empirical evidence suggests that preferences are not always well-defined. The first essay studies the descriptive and predictive power of the axiomatised expected utility theory and its alternatives, specifically, Tversky & Kahneman (1992)'s cumulative prospect theory and Bordalo, Gennaioli & Shleifer (2012)'s salience theory. We conduct a Lab experiment with binary choice questions over lotteries and find that both alternatives race closely and outperform expected utility. The second essay examines the economic importance of anchoring. Anchoring is proven to be robust in the psychology literature, but the quantitative economic significance of the phenomenon has not been given enough focus. We conduct a systematic synthesis of experiments examining the effects of numerical anchors on willingness to pay (WTP) and willingness to accept (WTA), and find that the effect of anchoring is relatively smaller than previously believed. Another key aspect of bounded rationality is the premise that people have limited computational power, and examining the economic and political implications of this is a fundamental task of modern social scientists. My third essay studies these implications in a voting environment with biased polls. In our experimental design, there is a strict subset of voters that is informed about the quality of the candidates, and polls serve to communicate this information to uninformed voters. Voters in the treatment group are presented with biased poll results, which favour systematically one candidate. The result shows that voters fail to infer the biased rules behind information revelation and account for it, since voters in the treatment group consistently elect the candidate favoured by polls more often than in the unbiased control conditions.

Declaration of Authorship

I, Lunzheng Li , declare that the thesis entitled *Essays in Bounded Rationality and Economic Experiments* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:.....

Date:.....

Co-Authorship Statement

Chapter 3 (*How Economically Important is Anchoring? A Research Synthesis of WTP/WTA Studies*) was co-written with Zacharias Maniadis (University of Southampton) and Constantine Sedikides (University of Southampton). My contribution to the production of these research works is outlined below:

- Literature searching and screening – Shared responsibility with co-author.
- Data collection and coding – Shared responsibility with co-author.
- Data analyses – Shared responsibility with co-author.
- Manuscript preparation – Shared responsibility with co-author.

Chapter 4 (*Can Biased Polls Distort Electoral Results? Evidence from the Lab and Field*) was co-written with Aristotelis Boukouras (University of Leicester), Will Jennings (University of Southampton) and Zacharias Maniadis (University of Southampton) . My contribution to the production of these research works is outlined below:

- Experimental design – Shared responsibility with co-author.
- Data collection – Shared responsibility with co-author.
- Data analyses – Shared responsibility with co-author.
- Manuscript preparation – Shared responsibility with co-author.

Acknowledgements

I would like to express my deepest gratitude to my main supervisor, Zacharias Maniadis, for his guidance, support and encouragement. I had the honour to read his PhD thesis, in which he thanked his adviser at UCLA and wrote:“ His consistent support, even from a distant location, is a great example for me to follow with my own students”. I will follow Zach's example just as he followed his adviser's. I thank my secondary supervisor, Michael Vlassopoulos, for numerous comments, suggestions throughout the PhD training, and for his support during the stressful job market period.

Special thanks to Zach for contributing to Chapter 3 and 4, and to Constantine Sedikides for contributing to Chapter 3, and to Aristotelis Boukouras and Will Jennings for contributing to Chapter 4. I also thank Jozef Bavolar, Magdalena Brzozowicz, Tore Ellingsen, Schlapfer Felix, Nathan Fong, Sarah Tanford for kindly sharing the raw data of their work, which is critical for Chapter 3. I thank the Economic and Social Science Research Council for its financial support.

I thank my friends and colleagues in the department, Chi Wan Cheang, Armine Ghazaryan, Xiaocheng Hu, Larissa da Silva Marioni, Abu Siddique, Marius Strittmatter, Tao Wang for helpful discussions and intellectual inspirations. I am also grateful to my best friends in China, Hao Yu and Wanlong Xu, for always being there for me in both my highs and lows.

I never thank my parents enough for their unconditional love.

Contents

1 Introduction **1**

2 Accounting for the Instability of Risk Preferences: Salience Theory versus Cumulative Prospect Theory **5**

2.1 Introduction . . . 5
2.2 Review of theories . . . 7
2.2.1 Expected Utility Theory . . . 7
2.2.2 Cumulative Prospect Theory . . . 7
2.2.3 Salience Theory . . . 9
2.3 The Experiment . . . 11
2.3.1 Stage 1 . . . 12
2.3.2 Stage 2 . . . 12
2.3.3 Implementation . . . 14
2.4 Results . . . 14
2.4.1 Calibrations . . . 14
2.4.2 Descriptive Power . . . 18
2.4.3 Predictive Power . . . 19
2.4.4 Additional Analysis . . . 21
2.5 Conclusion . . . 24

Appendix A **24**

A.1 Experimental Instruction . . . 25
A.2 Questions in Stage 2 . . . 27
A.3 Additional Tables and Graphs . . . 29

3 How Economically Important is Anchoring? A Research Synthesis of WTP/WTA Studies **31**

3.1 Introduction . . . 31
3.2 Methods . . . 33
3.2.1 Effect Sizes . . . 33
3.2.2 Literature Search and Inclusion Criteria . . . 34
3.2.3 Moderators . . . 35
3.3 Meta-Analytic Results . . . 39
3.3.1 Description of Studies . . . 39
3.3.2 Average Effect Size . . . 40
3.3.3 Moderator Analyses . . . 43
3.4 Robustness Checks . . . 45
3.4.1 Publication Bias . . . 45
3.4.2 Other Potential Biases . . . 47
3.5 Conclusion . . . 49

Appendix B **50**
B.1 Forest plots with z, complete dataset . . . 51
B.2 Meta-regression on r, complete dataset . . . 53
B.3 Meta-analytic results, reduced dataset with (Green, Jacowitz, Kahneman & McFadden, 1998) being excluded (52 studies) . . . 53
B.4 Descriptions and Meta-analytic results, reduced dataset (24 studies) . . . 57
B.5 Coding . . . 60

4 Can Biased Polls Distort Electoral Results? Evidence from the Lab and Field **63**
4.1 Introduction . . . 63
4.2 Related Literature . . . 66
4.3 Bias in Online Propagation of Poll Results: Evidence from the Field . . . 68
4.3.1 Opinion Polling in the US and UK . . . 68
4.3.2 Opinion Polling Data on Social Media (Twitter) . . . 69
4.4 Our Experimental Environment . . . 73
4.4.1 Voters' Preferences on Candidates, Voter Information, Polls, and Elections . . . 74
4.4.2 The Three Experiments . . . 77
4.5 Results and Descriptive Analysis . . . 79
4.5.1 Experiment 1 . . . 79
4.5.2 Experiment 2 . . . 80
4.5.3 Experiment 3 . . . 84
4.5.4 Additional Descriptive Analysis . . . 85
4.6 Regression Analysis . . . 90
4.7 Discussion and Conclusions . . . 98

Appendix C **100**
C.1 Experimental Instructions . . . 101
C.2 Additional Graphs . . . 104
C.3 Additional Tables . . . 113

Bibliography **115**

List of Figures

2.1 Distributions of estimates . . . 17
2.2 Kernel density estimation of AIC . . . 19
2.3 Kernel density estimation of predicted log-likelihoods . . . 20
2.4 Scatter of predicted log-likelihood against AIC . . . 21
2.5 Kernel density estimations of AIC and predicted log-likelihood (LCPT and NST) . . . 23
2.6 Distributions of the local thinking parameter: downside salient vs. upside salient . . . 24
A.1 Stage 1 sample screen . . . 26
A.2 Stage 2 sample screen . . . 26
A.3 Distributions of estimates for LCPT and NST . . . 30

3.1 Summary of number of studies in each category, complete dataset (53 studies) . . . 41
3.2 Number of studies against publication year and sample size, complete dataset (53 studies) . . . 41
3.3 Effect size as a function of publication year and sample size, complete dataset (53 studies) . . . 42
3.4 Funnel plot with z, complete dataset (53 studies) . . . 46
3.5 Residual plot with z, complete dataset (53 studies) . . . 47
3.6 Residual plot with z, reduced dataset (45 studies) . . . 48
B.1 Fixed effect model, complete dataset (53 studies) . . . 51
B.2 Random effect model, complete dataset (53 studies) . . . 52
B.3 Funnel plot with z, reduced dataset (52 studies) . . . 56
B.4 Residual plot with z, reduced dataset (52 studies) . . . 56
B.5 Summary of number of studies in each category, reduced dataset (24 studies) 57
B.6 Number of studies against publication year and sample size, reduced dataset (24 studies) . . . 57
B.7 Effect size as a function of publication year and sample size, reduced dataset (24 studies) . . . 58

4.1 Adjusted predictions (with 95% confidence intervals) of the number of retweets, by ΔVote . . . 73
4.2 Ideological preferences in the experimental interaction . . . 75
4.3 Sequence of actions in each experimental round . . . 76
4.4 Descriptive results of E1 . . . 81
4.5 Descriptive results of E2 . . . 83
4.6 Descriptive results of E3 . . . 86
4.7 Distribution of differences in the vote share of K: revealed poll results vs. elections . . . 99
C.1 Poll outcomes in treatment sessions of E1 (T1-T4) . . . 104

C.2 Average beliefs vs. poll outcomes in control sessions of E1 (C1-C4) 105
C.3 Average beliefs vs. poll outcomes in treatment sessions of E1 (T1-T4) . . 106
C.4 Poll outcomes in treatment sessions of E2 (T1-T4) 107
C.5 Average beliefs vs. poll outcomes in control sessions of E2 (C1-C4) 108
C.6 Average beliefs vs. poll outcomes in treatment sessions of E2 (T1-T4) . . 109
C.7 Poll outcomes in treatment sessions of E3 (T1-T5) 110
C.8 Average beliefs vs. poll outcomes in control sessions of E3 (C1-C4) 111
C.9 Average beliefs vs. poll outcomes in treatment sessions of E3 (T1-T5) . . 112

List of Tables

2.1 Possible states of the world in the example . . . 9
2.2 Descriptive Summary of Estimates . . . 15
2.3 Summary of AIC . . . 18
2.4 Ranking based on AIC . . . 19
2.5 Summary of predicted log-likelihoods . . . 20
2.6 Ranking based on predicted log-likelihoods . . . 20
2.7 Summary of AIC and predicted log-likelihoods (LCPT and NST) . . . 22
A.1 Descriptive Summary of Estimates (LCPT and NST) . . . 29

3.1 Summary of Moderators . . . 36
3.2 Summary of included articles . . . 40
3.3 Sub-group random-effect estimates of the overall ES, complete dataset (53 studies) . . . 43
3.4 Meta-regression on z, complete dataset (53 studies) . . . 45
3.5 Meta-regression on z, reduced dataset (45 studies) . . . 49
B.1 Meta-regression on r, complete dataset (53 studies) . . . 53
B.2 Sub-group random-effect estimates of the overall ES, reduced dataset (52 studies) . . . 54
B.3 Meta-regression on z, reduced dataset (52 studies) . . . 55
B.4 Meta-regression on z, reduced dataset (24 studies) . . . 59
B.5 Coding . . . 60

4.1 Twitter reporting of poll estimates in the US and the UK . . . 70
4.2 Selective propagation of poll estimates of the US 2016 presidential election 71
4.3 Selective propagation of poll estimates of voting intention, UK . . . 71
4.4 The experimental design . . . 78
4.5 Example presentation of poll results in each condition . . . 78
4.6 Number (percentage) of elections won for each party in each treatment and results of Fisher's exact test . . . 79
4.7 Average payoffs in each session . . . 89
4.8 Behaviour of informed voters . . . 89
4.9 Comparison of individuals' voting at the polls vs. the final election . . . 90
4.10 Effect on Beliefs (Model 1) . . . 92
4.11 Effect on Average Poll Information (Model 2) . . . 95
4.12 Effect on the Differences between Beliefs and Average Poll Information (Model 3) . . . 97
C.1 Descriptive summary of voting behaviour at the poll stage, pooled at session level, E1 . . . 113
C.2 Descriptive summary of voting behaviour at the poll stage, pooled at session level, E2 . . . 113

C.3 Descriptive summary of voting behaviour at the poll stage, pooled at session level, E3 113

Chapter 1

Introduction

Economic man "has complete, fully ordered preferences, perfect information and all the necessary computing power, and after deliberation, he or she chooses the action that satisfies their preferences better than any other" (Hargreaves-Heap & Clark, 2017). This is the central assumption of neoclassical economics. Such assumption has been under heavy critique since its emergence, and one of the most famous and sardonic critiques among them is made by Veblen (1898):

> *The hedonistic conception of man is that of a lightning calculator of pleasures and pains who oscillates like a homogeneous globule of desire of happiness under the impulse of stimuli that shift him about the area but leave him intact.*

Herbert A. Simon, probably not as harsh as Veblen, also criticised neoclassical economics for its tenet of perfect rationality. He coined the term 'bounded rationality' in his *Models of Man* (Simon, 1957), and stressed the cognitive limits and the processes of human reasoning. The present thesis studies bounded rationality in three different contexts: Chapter 2 studies expected utility theory and its alternatives in risk settings; Chapter 3 studies the effect of anchoring on the elicitation of economic evaluations; Chapter 4 studies information aggregation in a voting environment. The methodological tool of economic experiments is our chosen lens used to explore the relevant research questions. In particular, we conducted a series of lab experiments in Chapter 2 and Chapter 4. In Chapter 3, we performed a meta-analysis and synthesised experimental results in a literature that we wished to explore. An overview of the motivations, methods and findings of this thesis is presented below.

We started with an examination of expected utility theory and its alternatives. In the early years of economics, an economic agent was very much considered as, if we borrow Veblen's term, "a lightning calculator of pleasures and pains" and economic modelling was based on utility, which is the measure of "pleasures and pains". However, this sense of cardinal utility is no longer present in mainstream economics. Modern economic modelling is based on preferences structured through axioms and utility is no more than a representation of preferences. The landmark event in such evolution is Von Neumann & Morgenstern (1947)'s axiomatisation of expected utility theory. Expected utility is considered the standard representation of rational preference in choice under risk ever since.

It is well-known that expected utility has faced challenges as a descriptive theory of behaviour, and alternatives of expected utility theory are actively being developed and tested. In Chapter 2, "*Accounting for the Instability of Risk Preferences: Salience Theory versus Cumulative Prospect Theory*", we focus on a recently developed alternative – Bordalo et al. (2012)'s salience theory. The core psychological intuition behind this theory is that decision-makers have limited attention, and their attention is drawn to 'whatever is odd, different or unusual' (Kahneman, 2011). Bordalo et al. (2012) claimed that their theory provides a unified explanation for several intriguing phenomena related to the instability of risk attitudes, such as the Allais Paradox and excessive risk-seeking behaviours. Therefore, we designed choice questions for which the majority of subjects are likely to exhibit unstable risk attitudes, in order to investigate the descriptive and predictive power of this salience theory in the economic laboratory. We compared its performance with expected utility theory's and cumulative prospect theory's (Tversky & Kahneman, 1992). We found that salience theory and cumulative prospect theory outperform expected utility theory, which does not account for the instability of risk preferences. Moreover, cumulative prospect theory outperforms salience theory by an insignificant margin. We attribute this small gap to the unsophisticated specification of the salience function and the substantial heterogeneity of the 'local thinking' parameter. We argue that salience theory captures important features of unstable risk preferences, yet further work on the functional representation of the theory is necessary to make it as applicable as cumulative prospect theory.

Then, we look further into the assumption of 'complete, fully ordered preferences'. Standard economic theory suggests that preferences are well-defined, in the sense that they can be represented by pre-defined utility functions. However, empirical evidence, such as preference reversals (Lichtenstein & Slovic, 1971; Grether & Plott, 1979) and the framing effect (Kahneman & Tversky, 1981), show that preferences are "often ill-defined, highly malleable and dependent on the context in which they are elicited" (Camerer & Loewenstein, 2003, p. 15). In Chapter 3, "*How Economically Important is Anchoring? A Research Synthesis of WTP/WTA Studies*", we focus on one particular phenomenon that reflects such ill-defined preferences, namely anchoring. Anchoring can be defined as the influence of a normatively irrelevant cue on a subsequent expression of judgement, and it is considered one of the most robust psychological phenomena in judgement and decision-making. Early literature (Northcraft & Neale, 1987; Green et al., 1998; Ariely, Loewenstein & Prelec, 2003) shows that anchoring is relevant for the elicitation of economic preferences, in a strong and robust manner. However, subsequent studies (Fudenberg, Levine & Maniadis, 2012; Maniadis, Tufano & List, 2014) found weaker and less robust effects. To examine the quantitative economic significance of anchoring, we explored the experiments in literature and performed a systematic synthesis of relevant studies.

We include 53 studies from 24 articles and choose the Pearson correlation coefficient between the anchor number and target response (in our case, WTP/WTA) as the primary effect size. Both fixed-effects and random-effects models point to a moderate

overall effect (less then 0.3), which is smaller than early influential studies. Further meta-regression analysis shows that subjects in WTP tasks are more likely to be influenced, comparing to subjects in WTA. Incentives do not attenuate the effects. The relevance and compatibility of the anchor to the target response and the experiment type also matter for the magnitude of the effect size. Overall, the effect of anchoring on economic evaluation should not be overlooked, but it does not seem to be as strong as previously believed.

In the final chapter, "*Can Biased Polls Distort Electoral Results? Evidence from the Lab and Field*", we examine the political implications of the fact that people often have limited ability to make proper inferences and to learn from the evidence, especially if this evidence is biased. In particular, we examine the issue in the context of a two-candidate voting environment to see if voters can make proper inferences using the results of biased pre-election polls and the results of previous electoral results. We first show empirically how modern communication (through social media) may naturally result in such biased exposure to pre-election polls. Then, in a series of experiments with a total of 375 participants, we investigate the impact of such biased exposure on election outcomes. In our design, a subset of voters has information on the quality of two candidates, while the remaining voters are uninformed. Thus, polls serve to communicate information to uninformed voters. In our control group, participants have access to the set of all polls available (unbiased polls), whereas in the treatment group, participants observe only the polls most favourable to one candidate (biased polls). We find that biased polls consistently provide an electoral advantage to the party that was favoured by the bias. Remarkably, this holds even when voters are a priori informed about the bias. Also, the results show limited evidence that the repeated opportunities for learning allowed voters to understand the systematic bias and account for it. We also argue that this limited ability of inferring and learning in the lab can be generalised to real elections, as real elections are more complicated.

Chapter 2

Accounting for the Instability of Risk Preferences: Salience Theory versus Cumulative Prospect Theory

2.1 Introduction

Expected utility theory (henceforth denoted as EUT), the standard tool with which economists used to model risk, assumes that individuals respond to risk in a consistent manner. The theory may provide a valuable normative guide, but it faces difficulties while playing the role of a descriptive theory. Mounting empirical evidence suggests that people systematically switch between risk aversion and risk-seeking, depending on the situation. Allais (1953) shows that when people choose between two lotteries, adding a common consequence to both lotteries might change the preference order, which contradicts the independence axiom. Lichtenstein & Slovic (1971) find that subjects tend to choose relative safer lotteries when making choices, but are willing to pay more for risky ones. Moreover, Kahneman & Tversky (1979, 1981) suggest that framing lotteries differently or replacing gains with losses could lead to a reversal of preference orders and other paradoxes.

In the past several decades, alternatives of EUT have been developed to explain this instability of risk preference. In this paper, we are particularly interested in Bordalo et al. (2012)'s salience theory (henceforth denoted as ST). The idea of the theory is that a decision maker's attention is drawn to salient consequences, and the probabilities are distorted accordingly. Bordalo et al. (2012) introduced the theory as a unified explanation for several anomalies related to unstable risk preferences, such as excessive risk-seeking behaviour and the Allais paradox. Importantly, the new theory can explain these phenomena using only a small set of assumptions about the function that guides

how salient lottery outcomes are perceived. ST also makes new predictions which contradict prospect theory (Kahneman & Tversky, 1979). In addtion, the authors presented the results of a series of online experiments as empirical evidence.

This novel theory has attracted considerable attention from empirical reseachers. Kontek (2016) points out that the certainty equivalents of lotteries are undefined according to the theory for some range of probabilities, and the theory also violates monotonicity. Nielsen, Sebald & Sørensen (2018) report an online experiment with 473 participants. They manipulate the salience value of the good and bad consequences in each lottery and the results are consistent with the prediction of ST. Frydman & Mormann (2018) replicate the Allais paradox experiment with one adjustment: they set two lotteries to be correlated. ST relies on the joint distribution of the lotteries, while other theories, such as Tversky & Kahneman (1992)'s cumulative prospect theory (henceforth denoted as CPT) do not, so Frydman & Mormann (2018) conclude that ST provides a coherent framework to understand the Allais paradox. Dertwinkel-Kalt & Köster (2019) link risk preference to the skewness of the probability distribution of lotteries, and experimentally show that ST accommodates such preferences better than CPT. Königsheim, Lukas & Nöth (2019) focus on the local thinking parameter, which measures how much individuals' decision weights are distorted because of limited attention. They calibrate the parameter and argue that the estimate depends on whether the lottery is downside or upside salient.[1]

We report a laboratory experiment conducted to examine the empirical validity of ST. The experiments mentioned above test ST on the basis of its axiomatic fundamental and behavioural assumptions, i.e., the researchers examine whether salience affects risk taking or not and how. However, our experiment is a different exercise. The purpose of this paper is not to verify or challenge the behavioural tenets of ST, but to compare the theory to other candidate theories in the sense of empirical fitness and predictions. We apply the theory in a straightforward manner to a risk choice setting where the majority of subjects exhibit unstable risk attitudes and check how accurately the theory describes and predicts decisions. We choose the popular CPT as the baseline model for evaluation since this theory has been tested thoroughly and it is useful in terms of application (Gonzalez & Wu, 1999; Wu & Markle, 2008; Hey, Lotito & Maffioletti, 2010; Kothiyal, Spinu & Wakker, 2014; Georgalos, 2019). The classical EUT is also included in the final comparison. The results show that in terms of both descriptive and predictive power, CPT outperforms ST and ST outperforms EUT. It is to be expected that EUT be dominated since it does not account for unstable risk preferences. On the other hand, CPT and ST are racing closely in terms of predictive power. In future research, the gap may be reduced by improving two aspects of ST: (i) The functional representation of salience function; (ii) Deriving different local thinking parameters for lotteries with opposite salience directions.

[1] We examined the local thinking parameter in the same manner. See Section 2.4.4 for details.

The structure of the paper is as follows. In Section 2.2, we review the investigated models and introduce the functional forms used in the analysis. Section 2.3 introduces the experimental design. The results are presented in Section 2.4. Section 2.5 concludes.

2.2 Review of theories

The theories under investigation are EUT, CPT and ST. In this section, we review the theories and present the preference functional used in our analysis. We also explain how the theories can (or cannot) account for phenomena related to unstable risk preferences. Kahneman & Tversky (1979)'s version of common consequence effect example (Equation 2.1) is used for illustration. In their experiment, subjects are asked to choose between $L(c)$ and $R(c)$ for different values of c. It is clear that $R(c)$ is the safer option for any c. The results show that the majority of subjects choose the safer option $R(c)$ when $c = 2400$. However, most of them switch to the riskier option $L(c)$ when $c = 0$.

$$L(c) = \begin{cases} 2500, & \text{with prob.} \quad 0.33 \\ 0, & \qquad\quad 0.01 \\ c, & \qquad\quad 0.66 \end{cases} \quad ; \quad R(c) = \begin{cases} 2400, & \text{with prob.} \quad 0.34 \\ c, & \qquad\quad 0.66 \end{cases} \tag{2.1}$$

2.2.1 Expected Utility Theory

An EUT agent's preference order over $L(c)$ and $R(c)$ does not depend on the value of c because of the independence axiom, and her risk preference can be characterised simply by the curvature of the utility function. Our chosen utility function is:

$$U(x) = x^{\tau} \tag{2.2}$$

where $\tau > 0$. Note that $0 < \tau < 1$ indicates risk-averse, $\tau = 1$ indicates risk neutral, and $\tau > 1$ indicates risk-seeking preferences.

2.2.2 Cumulative Prospect Theory

CPT is an advanced version of prospect theory. In the basic expected utility framework, a subject is assumed to have an underlying value function and the utility is linear in probabilities. Prospect theory maintains the idea of the fixed value function, but the value is relative to a reference point. However, instead of using raw probabilities as decision weights, the theory converts the probabilities into decision weights according to a non-linear weighting function. Kahneman & Tversky (1979) derive the weighting function based on psychological insights and the function overweights small probabilities and underweights moderate and high probabilities. In terms of the Kahneman & Tversky (1979)'s common consequence effect example, when c switches from 0 to 2400, the probability associated with the zero payoff in $L(c)$ drops significantly (from 0.67 to

0.01). Thus the probability is being overweighted, and subjects choose the safer option $R(c)$.

The shape of the non-linear weighting function plays a vital role in explaining the common consequence effect and other anomalies. However, a mere monotonic transformation of outcome probabilities causes violations of stochastic dominance and the problem of non-additivity.[2] To solve these problems, Tversky & Kahneman (1992) modified the weighting strategy by incorporating rank-dependent utility (Quiggin, 1982): Instead of transforming each probability separately, the new model transforms the entire cumulative distribution function. Considering a risky prospect which is represented as n pairs of (x_i, p_i), where x_i is the payoff and p_i is the corresponding probability, and $x_1 < x_2 < ... < x_n$, the decision weight π_i equals $w(p_i + ... + p_n) - w(p_{i+1} + ... + p_n)$. $w(.)$ is the cumulative probability weighting function, and $w(p_i + ... + p_n)$ measures the probability of getting a value at least as good as x_i. On the other hand, the features of the value function are maintained. The value function is defined on deviations from a reference point, is concave for gains and convex for losses and it is steeper for losses than for gains.

We follow the original parametric representation of the cumulative probability function and value function of Tversky & Kahneman (1992). Our experiment does not deal with losses,[3] thus the weighting function $w(p)$ is:

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}} \tag{2.3}$$

where $0 < \gamma < 1$. The value function is:

$$v(x) = x^\alpha \tag{2.4}$$

where $0 < \alpha < 1$, and we assume that the reference point for evaluation is zero.

[2] Consider the following example. Lottery A has two outcomes: 1 and 100, and the corresponding probabilities are 0.01 and 0.99. Lottery B's outcomes are all the integers from 1 to 100, and each outcome corresponds to a probability of 0.01. Obviously, lottery A first-order stochastically dominates lottery B. However, violations of stochastic dominance may occur if the outcome probabilities are distorted in the manner of a monotonic weighting function (low probabilities get larger while high probabilities get smaller). Moreover, for lottery B, the weights obviously do not add to unity after the distortion.

[3] We choose not to deal with losses for two reasons: Firstly, ST and CPT have the same degrees of freedom if negative payments are not included. This makes the comparison simpler and more meaningful. Secondly, including losses increases the decisions subjects need to make and inevitably increases the experiment time, which was a constraint in our case.

2.2.3 Salience Theory

Table 2.1: Possible states of the world in the example

States	Payoff combination	Probability
s_1	(x_1^1, x_2^1)	$\pi_{s_1} = p_1 p_2$
s_2	(x_1^1, x_2^2)	$\pi_{s_2} = p_1(1-p_2)$
s_3	(x_1^2, x_2^1)	$\pi_{s_3} = (1-p_1)p_2$
s_4	(x_1^2, x_2^2)	$\pi_{s_4} = (1-p_1)(1-p_2)$

Note: There are two lotteries: $L_1 = (x_1^1, p_1; x_1^2, 1-p_1)$ and $L_2 = (x_2^1, p_2; x_2^2, 1-p_2)$.

Unlike CPT or several other generalisations of EUT which focus on the shape of the probability weighting function, the key point of ST is that the distortion on probabilities depends on the payoffs of both lotteries, specifically, on "how salient a state is". Bordalo et al. (2012) refer to decision makers as local thinkers, i.e., they can only think "locally" due to limited attention or cognitive limitations. Therefore, a local thinker can only focus on (or process) the most salient state, and tends to overweight it. We present the model with a two-lottery choice set $\{L_1, L_2\}$ where both lotteries have two possible outcomes, i.e., $L_i = (x_i^1, p_i; x_i^2, 1-p_i)$ where x_i^1, x_i^2 are the possible outcomes, $p_i, 1-p_i$ are the corresponding probabilities, and $i \in \{1, 2\}$. This choice problem can be described as a set of states of the world $S = \{s_1, s_2, s_3, s_4\}$, and state $s \in S$ has a probability of π_s (see Table 2.1).

A salience function $\sigma(.)$ is defined over the payoff combinations and is used to measure the perceived difference between the two payoffs in each state. It satisfies two conditions: ordering (two payoffs define an interval for each state, and a state is less salient if the corresponding interval is a subset of the alternative) and diminishing sensitivity (keeping the payoff difference constant, a state is less salient when payoffs lies futher from zero).[4] Bordalo et al. (2012) suggest the following continuous and bounded function for state s:

$$\sigma(x_i^s, x_j^s) = \frac{\left|x_i^s - x_j^s\right|}{|x_i^s| + \left|x_j^s\right| + \beta} \quad , (\beta > 0). \tag{2.5}$$

The local thinker ranks the states according to the value of the salience function, with lower k_s (k_s is a positive integer and it represents the ranking of state s) indicating higher salience, and the distorted decision weight is given by:

$$\pi_s^d = \pi_s \times \frac{\delta^{k_s}}{\sum_{r \in S} \delta^{k_r} \pi_r} \tag{2.6}$$

[4] Bordalo et al. (2012) also mentioned a "reflection condition" to extend the theory to losses, which is not within the scope of our study.

where $0 < \delta \leq 1$, and δ is the local thinking parameter which measures how much a local thinker's attention is drawn by the salient states. $\sum_r \delta^{k_r} \pi_r$ is used to normalize $\sum \pi_s^d$ to 1. Therefore, states ranked higher (the most salient states) are overweighted and states ranked lower (the least salient states) are under-weighted. Bordalo et al. (2012) assume a linear value function $v(x) = x$, and the local thinker evaluates L_i as:

$$V(L_i) = \sum_{s \in S} \pi_s^d x_i^s \tag{2.7}$$

Consider Kahneman & Tversky (1979)'s common consequence effect example. When the common consequence c is 0, the most salient state is associated with the payoff combination (2500, 0), which makes the riskier $L(c)$ more attractive. However, when the common consequence c becomes 2400, the most salient state becomes (0, 2400) where the payoff of $R(c)$ clearly dominates. Thus, subjects choose the safer option $R(c)$.

Several auxiliary assumptions on the basic ST framework are necessary in order to obtain reasonable parameter estimates. The ranking k_s in Equation 2.6 creates a discontinuity in the utility function, which gives us difficulty in estimating the salience function and it does not contain the information of the magnitude of salience in distortions. Since a more salient state is associated with a smaller k_s, we replace k_s with $\frac{1}{\sigma(x_i^s, x_j^s)}$ to smooth out the utility function.[5] Also, to avoid the denominator becoming zero, we modify the salience function to the following:

$$\sigma(x_i^s, x_j^s) = \frac{\left|x_i^s - x_j^s\right| + \theta}{|x_i^s| + \left|x_j^s\right| + \beta} \quad , (\theta, \beta > 0), \tag{2.8}$$

without loss of generality, we set $\beta = \lambda\theta$ with $\lambda > 0$. This salience function satisfies the ordering condition indicating that $\sigma(x_i^s, x_j^s)$ increases with the gap between x_i^s and x_j^s. Therefore, for any fixed x_j^s, the function is increasing in x_i^s. In a special case when $x_j^s = 0$ and $x_i^s > 0$, we have:

$$\sigma(x_i^s, 0) = \frac{x_i^s + \theta}{x_i^s + \lambda\theta} = 1 + \frac{(1-\lambda)\theta}{x_i^s + \lambda\theta} \quad , (\lambda, \theta > 0), \tag{2.9}$$

to satisfy the ordering condition, λ should be larger than 1. We set λ to e (Euler's Number) for calibration purposes.[6] We use the following salience representation in the analysis:

$$\sigma(x_i^s, x_j^s) = \frac{\left|x_i^s - x_j^s\right| + \theta}{|x_i^s| + \left|x_j^s\right| + e\theta} \quad , (\theta > 0), \tag{2.10}$$

[5] Following Bordalo et al. (2012)'s suggestion, we replaced k_s with $-\sigma(x_i^s, x_j^s)$ at first. However, this approach is inappropriate for our case since it yields extreme estimates (both δ and β are extremely small).

[6] If λ is a rational number, for any payoff combinations which satisfy $\frac{|x_i^s - x_j^s|}{|x_i^s| + |x_j^s|} = \frac{1}{\lambda}$, $\sigma(x_i^s, x_j^s)$ is a constant which clearly violates diminishing sensitivity. The irrational number e is chosen only for its aesthetic feature and it is mathematically irrelevant.

In general, this setting satisfies the conditions of the theory and gives us reasonable estimates of θ and δ. Regarding the value function, we stick to the linear value so that CPT and ST have the same number of parameters.

2.3 The Experiment

We use a series of binary choice questions to elicit risk preferences. In the experimental practice of risk preferences elicitation, researchers usually ask subjects three forms of questions: binary choice questions, reservation price questions and allocation questions (Hey & Pace, 2014).[7] Allocation questions are not our choice because of the context-dependence nature of ST. In particular, an ST agent judges a lottery differently when the alternative is different, hence the theory does not have a unified preference functional form for a single lottery, which makes the allocation question method infeasible. Accordingly, the other two forms of questions have special advantages. According to Hey et al. (2010), binary choice questions "are easier to explain to subjects; easier for them to understand; and less prone to problems of understanding associated with the various mechanisms for eliciting", while obtaining the reservation prices (certainty equivalents) of lotteries enhances the information in data. We believe that the attractions of both methods are important to our experiment. Therefore, in the first stage, we let subjects make choices between a lottery and a set of consecutive sure payoffs, so that the certainty equivalents of the lottery can be estimated if necessary.[8] In the second stage, a subject is given a series of binary choice questions and she is paid according to her decision of one randomly chosen question. The observations derived in the two stages are used for different purposes, which we shall explain later.

As mentioned above, we shall judge a theory based on its descriptive and predictive power rather than its behavioural hypothesis. We follow Hey et al. (2010) in determining which theory is 'better'. Briefly speaking, subjects are required to answer two sets of binary choice questions, with part of the observations used for calibration, and the remaining part used for predictive capacity testing. In particular, predictive capacity can be measured by predicted log-likelihoods, and the theory which has larger predicted log-likelihoods is considered to be outperforming the others. Also, the comparison is not based on statistical tests, as according to (Hey et al., 2010, pp. 83), "statistical significance tells us nothing about economic significance".

We design our experiment in two stages: Stage 1 is used for calibration and testing descriptive power, and Stage 2 is used for testing predictive capacity. Stage 1 questions are designed to give enough information for calibration. Stage 2 questions focus on two typical phenomena which exhibit unstable risk attitudes: risk-seeking behaviour (typically a risk-averter might prefer a relatively small chance of winning a big prize to

[7] A typical allocation question design is like the following: subjects are given a fixed amount of tokens and they are asked to allocate the tokens to events with different probabilities. Subjects maximise their preference functions to make the allocation decisions.

[8] Subjects' certainty equivalents or reservation prices are not elicited directly. However, this design gives us enough information to infer them.

the expected value of such lottery) and the common consequence effect (the preference order over two lotteries is affected by the change in a common consequence).

2.3.1 Stage 1

In Stage 1, each subject faces 36 rounds, and in each round, she needs to make decisions between a risky prospect and a sure outcome eight times. This design allows us to get as much information as we would by asking subjects for their reservation prices. In a typical round, the screen displays a lottery on the left and a descending series of eight sure outcomes on the right side, which is linearly spaced between a value £0.5 higher than the low outcome in the lottery and a value £0.5 lower than the high outcome (see Figure A.1 for a sample screen-shot). The basic design consisted of four two-outcome gambles crossed with nine probabilities associated with the high outcome. The four gambles are (in pounds) (50 - 0), (20 - 0), (15 - 5), (20 - 10). The nine probabilities are .01, .05, .10, .25, .50, .75, .90, 95, and .99. We have in total 36 prospects and hence 36 rounds in Stage 1. The order of the pages is randomised for each subject.

2.3.2 Stage 2

Two types of choice questions are used in Stage 2 to examine the mentioned risk-seeking behaviour and the common consequence effect: the 'mean-preserving' question (a choice between a sure payoff and its mean-preserving spread) and the common consequence effect question (a choice between two lotteries which share a common consequence). After a series of pilots, we chose to have 40 questions in Stage 2 because of budget and time constraints. We decided to use 20 'mean-preserving' questions and 20 common consequence effect questions, with every subject answering the same set of questions with different and randomised orders.[9]

The 'mean-preserving' question takes the following form:

$$A = \begin{cases} x + \frac{(1-p)a}{p}, & p \\ x - a, & 1-p \end{cases} \quad ; \quad B = \Big\{ x, \qquad 1 \tag{2.11}$$

where the expected values of A and B are equal. We firstly set $x \in \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$, and let a be any positive integer smaller than x. p is set to be a number in $\{0.25, 0.4, 0.5, 0.6, 0.8\}$ for two reasons: First, we pay subjects with real money, so payments with decimal positions less than two are preferred. Therefore, probabilities like 0.33 and 0.66 are not included; Second, we do not consider small p such as 0.01 and 0.05, since small probabilities would make the associated payoffs very large, and we constrain the largest payment in Stage 2 to be £25.[10] This leaves us with 5500

[9] See Appendix A.2 for questions used in real sessions.

[10] Subjects can usually finish Stage 2 in less than 30 minutes, and a possible payment of £25 is sufficiently large.

combinations.[11] After discarding questions with excessively high payments and those that are not payable (having more than two decimal points), 159 questions are left, and 20 of them are picked randomly for Stage 2. Only 20 questions are selected out of 5500, yet we argue that the selection is to some extent representative. After cutting off around 5000 combinations, the remaining still cover a wide range of payoffs. For instance, the sure payoff (x) ranges in {5, 6, 7, 9, 10, 11, 12, 13}, and the higher payoff ($x + \frac{(1-p)a}{p}$) in the lottery ranges in {6.5, 7.0, 8.0, 8.5, 9.0, 10.0, 11.0, 11.5, 12.0, 13.0, 14.5, 15.0}. Also, a potential drawback of using a comprehensive list of questions is that participants are likely to notice the pattern of the questions even if the order is randomised.

The common consequence effect question takes the following form:

$$C = \begin{cases} h, & p_h \\ z, & p_z \\ 0, & p_0 \end{cases} \quad ; \quad D = \begin{cases} l, & p_h + p_0 \\ z, & p_z \end{cases} \tag{2.12}$$

where $h > l$ and $z \in [0, l]$. p_h, p_z, and p_0 are the corresponding probabilities for payoffs h, z and 0 respectively. The probability corresponding to l is $p_h + p_0$. This form of pairwise choice can be seen as a generalised form of Tversky & Kahneman (1992)'s common consequence effect example. The consistently observed behaviour is that subjects shift from preferring the riskier option to preferring the safer one when z changes from 0 to l, i.e., when $z = 0$, $C \succ D$, and when $z = l$, $C \prec D$. Also, we believe that this behaviour is more likely to be triggered if h is only slightly larger than l and p_0 is relatively small.[12] Therefore, considering the budget, we set $h \in$ {10, 11, 12, 13, 14, 15}, and $l = h - a$ where $a \in$ {1, 2}, and $p_0 \in$ {0.01, 0.05}. Also, we let $p_h \in$ {0.1, 0.33, 0.5, 0.75, 0.9} and $z \in \{0, l/4, l/2, 3l/4, l\}$. This leaves us with 600 combinations, and then we eliminate combinations for which the two theories yield the same predictions conditional on the pilot calibrations ($\alpha = 0.86$,$\gamma = 0.62$, $\theta = 3.28$ and $\delta = 0.67$).[13] 150 questions are left and we randomly choose 20 questions for the remaining part of Stage 2.

[11] For any $x = k$, we have $a \in$ {1, 2, ..., k-2, k-1}. There are five possible probabilities, and $x \in$ {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}. Therefore, the total number of combinations is $5 \times \sum_{k=5}^{15} k(k-1) = 5 \times (\sum_{k=5}^{15} k^2 - \sum_{k=5}^{15} k) = 5 \times (\sum_{k=1}^{15} k^2 - \sum_{k=1}^{4} k^2 - \sum_{k=5}^{15} k) = 5500$.

[12] Allais (1953)'s original example satisfies a similar form, and simple intuition also supports the idea. For instance, if h is far greater than l, it is likely that C is preferred to D when $z = l$, and if p_0 is very large, subjects would preferred D to C when z is 0 as they would not risk a huge chance (p_0) of getting l for a slightly larger outcome h.

[13] In terms of the anomalies which the two theories can explain, CPT and ST still overlap. If the purpose is to 'race' between two theories, examining the overlapping part is not very meaningful. We shall try to focus on questions for which the two theories yield different predictions. Also, please note that the pilot sessions were conducted at the University of Southampton with 42 subjects.

2.3.3 Implementation

Our study has a sample size of 48 and subjects are students from the University of Southampton crossing all disciplines, 47% of them being females.[14] The experiment was conducted in the Social Sciences Experimental Laboratory (SSEL) at the University of Southampton using oTree (Chen, Schonger & Wickens, 2016). The average payment per subject was £16.07 and each session of the experiment lasted around 60 minutes including the payment stage. The payment includes a participation fee of £4, a fixed Stage 1 fee of £4 and the payment for a randomly chosen question of Stage 2 (the payment was determined by a ten-sided die, see details in Appendix A.1).[15] At the beginning of each session, subjects read the instructions, and the experimenters show them the die which will be used to determine their payoffs at the end of the experiment. Experimenters try to ensure that subjects trust the instructions, and that their payments only depend on their decisions and luck. There is no time constraint in answering the questions, and subjects do not need to wait for others to start the next question or stage. Some subjects may finish before others, and the following statement is displayed on their screens: "We are waiting for everyone to finish. Thank you so much for your patience". After everyone is finished, a random question is selected by the computer for each subject and is displayed on the screen with the subject's decision. The experimenters then go to the subjects one by one to roll the dice and record the payment. Subjects collect their payments and they are free to leave.

2.4 Results

2.4.1 Calibrations

The maximum likelihood method is used to estimate the parameters and obtain the predicted log-likelihoods, and assumptions need to be made regarding the stochastic nature of the data. Assume that a subject faces n binary choice questions: (L_1,R_1), $(L_2,$ $R_2)$, $(L_3,\ R_3)$,...,$(L_n,\ R_n)$, and the preference function is $V(.)$. Then for an arbitrary question i, in the absence of noise, the subject chooses $L_i\,(R_i)$ if $V_{L_i} > V_{R_i}\,(V_{R_i} > V_{L_i})$. With choice error ϵ, the subject prefers $L_i\,(R_i)$ to $R_i\,(L_i)$ only if $V_{L_i} - V_{R_i} + \epsilon > 0\,(V_{R_i} -$

[14] To determine the sample size, we refereed to studies such as Hey et al. (2010) and Hey & Pace (2014), which do not carry out statistical test in comparing theories. In particular, Hey et al. (2010) has a sample size of 48 and Hey & Pace (2014) recruits 129 subjects and put them into two treatments (around 65 subjects in each treatment). We recruited 49 subjects using ORSEE (Greiner, 2015) and exclude one subject because of extreme calibrations. This subject has $\tau = 2.75^{-10}$, $s_{eut} = 6.88^{-9}$, $\gamma = 0$, $s_{cpt} = 0$, and $\delta = 8.04^{-14}$ (see Section 2.4.1 for the notations). A close look at the data reveals that this subject is extremely risk-averse and they always chose the sure payoff in Stage 1 (for all 288 decisions). This is why τ, γ, and δ are extremely small, and the zero standard deviation of the error term makes it impossible to calculate the out-of-sample likelihoods. Also, it is clear that for such subject, the strategy in Stage 2 would be simply choosing the safer option. It is indeed the case for the subject, and out of 40 questions, they selected the safer option 38 times.

[15] We pay subjects a fixed amount in Stage 1 since a reasonable method (such as the Becker–DeGroot–Marschak method) to incentivise the subjects would increase the experiment time significantly, and the calibrations in our real sessions are comparable to the literature (see Section 2.4.1).

$V_{L_i} + \epsilon > 0$). Following the Fechner error specification,[16] we assume that the error term ϵ is normally distributed with mean zero and variance s^2. Let $\mathcal{L}_i$ denote the likelihood function of question i, and we have $\mathcal{L}_i = Prob(\epsilon > V_{R_i} - V_{L_i})$. Since $\epsilon \sim \mathcal{N}(0, s^2)$, $\mathcal{L}_i$ equals $1 - \Phi[(V_{R_i} - V_{L_i})/s]$ where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution, and the log-likelihood function is written as follows:

$$\sum_{i=1}^{n} ln(1 - \Phi[(V_{R_i} - V_{L_i})/s]). \tag{2.13}$$

Table 2.2: Descriptive Summary of Estimates

Theory	Parameter	Mean	Median	s.d.
EUT	τ	0.71	0.69	0.28
	s_{eut}	3.34	1.67	7.22
CPT	α	0.80	0.81	0.18
	γ	0.66	0.64	0.19
	s_{cpt}	2.42	1.61	3.51
ST	θ	29.03	1.23	106.76
	δ	0.67	0.72	0.30
	s_{st}	6.78	5.47	4.70

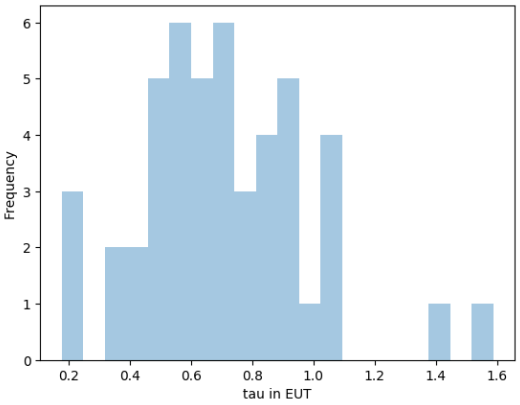
Note: The results exclude one subject for the reasons explained in footnote 14.

Observations of Stage 1 are used to pinpoint the functional form. The calibrations are at the individual level, and the summary of the results is presented in Table 2.2. The calibration result for EUT shows that the majority of subjects are risk-averse (42 out of 48), and for most of them, τ ranges from 0.5 to 0.9 (see Figure 2.1a). In terms of α and γ in CPT, the distributions of the values are depicted in Figures 2.1b and 2.1c. Comparing Figure 2.1a to Figure 2.1b, we see that incorporating non-linear probabilities affects the curvature of the value function significantly. Also, the fact that a high fraction of subjects has α near 1 may indicate that the S-shaped probability weighting function is crucial in describing the decision patterns. The calibrations are similar to the pilot sessions in which the medians of α and γ are 0.86 and 0.62 respectively. Besides, they are comparable to the original estimations made by Tversky & Kahneman (1992) (the medians of α and γ are 0.88 and 0.61). Regarding the calibrations of ST, δ is comparable to the pilot session calibrations (a median of 0.67) and Bordalo et al. (2012)'s calibration (0.7). Additionally, it is roughly consistent with Königsheim et al. (2019)'s result which shows that δ is between 0.7 and 0.8. However, according to Table 2.2, the estimates of θ show significant heterogeneity among individuals. Figure 2.1d depicts the distribution of θ, which shows that most estimates are smaller than 20. To be more specific, for 39 among 48 subjects, θ is smaller than 10. Considering only those 39 subjects, the

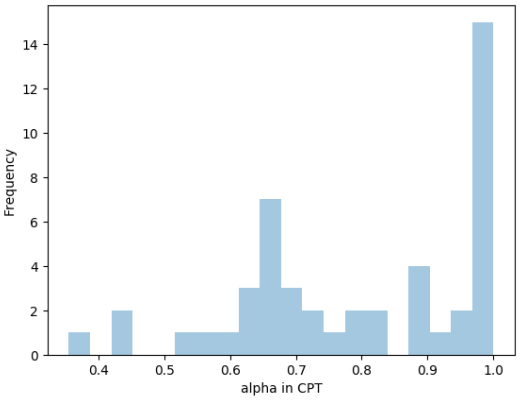
[16] We use the Fechner error story in our analysis, since it is relatively simple and most commonly applied in the relevant literature.

mean and median become 1.45 and 0.51, and the s.d. reduces to 1.87. Therefore, our estimates of θ are in fact not decentralised. Outliers with extremely large θ (greater than 400) magnify the standard deviation. We include the outliers into our analysis, since they do not affect the out-of-sample log-likelihoods. However, we acknowledge that our functional form of the salience function is not optimal. Optimising the form of the salience function is the topic of future research. In general, we obtain reasonable estimates of the parameters, and this is important as the result is sensitive to calibrations.

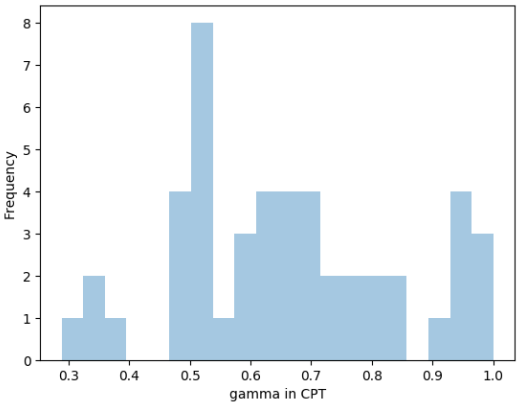
Figure 2.1: Distributions of estimates



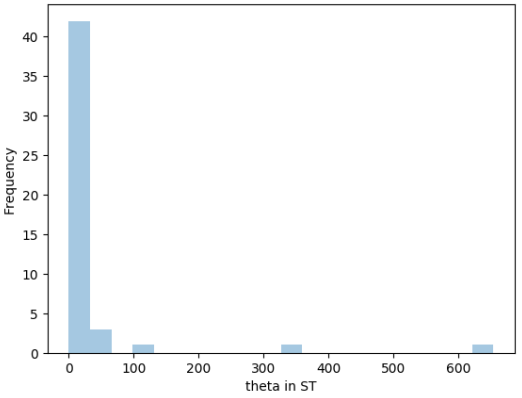
(a)



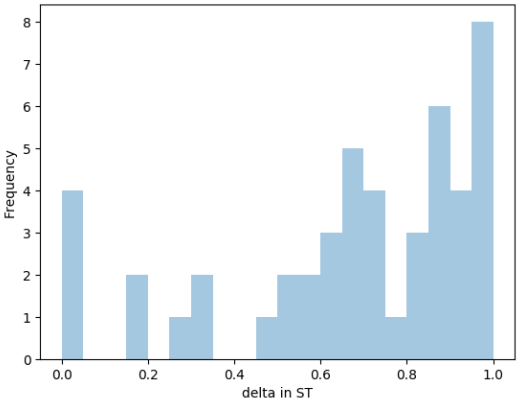
(b)



(c)



(d)



(e)

2.4.2 Descriptive Power

The measure of descriptive power is based on the fitted values of the maximised log-likelihoods in Stage 1. The fitted log-likelihoods are not compared directly since EUT has a different number of parameters relative to CPT and ST. To take into account the different degrees of freedom of the compared models, we apply the Akaike Information Criterion (Akaike, 1973)[17] which takes the following form:

$$AIC = 2k - 2ln(\mathcal{L}), \tag{2.14}$$

where k denotes the degrees of freedom, and $ln(\mathcal{L})$ represents the log-likelihood function. We firstly compare the mean, the median as well as the 5% and 10% trimmed means of AIC values across all subjects, and we find that CPT, in general, has the best fit and ST outperforms EUT. The results are reported in Table 2.3 (a smaller AIC value corresponds to a better fit). Besides, Akaike weights (Burnham & Anderson, 2002) are calculated using CPT as the benchmark model.[18] The average Akaike weights for EUT, CPT and ST are 14.97%, 62.53% and 22.5% respectively. This indicates that the probability of CPT being the best fitting model among the candidates is more than 60%, and ST has a slightly better chance than EUT. Also, we analyse the data at the individual level, and find the best fitted model for each subject. We report the percentages of subjects for which a given theory yields the best, second best, and worst fit in Table 2.4. The '1st' column indicates that for 56.25% of subjects CPT performs best, and the percentages for EUT and ST are 18.75% and 25% (comparable to the Akaike weights). Further, Figure 2.2 shows the Kernel density estimation of AIC for the three theories and it confirms the notion that CPT outperforms ST, and ST outperforms EUT in terms of descriptive adequacy.

Table 2.3: Summary of AIC

Theory	Mean	$\text{Mean}_{0.05}$	$\text{Mean}_{0.1}$	Median	s.d.
EUT	202.62	202.67	202.72	200.04	67.42
CPT	165.85*	164.56*	163.84*	153.07*	70.21
ST	188.01	190.03	186.95	186.28	77.40

Note: $\text{Mean}_{0.05}$ and $\text{Mean}_{0.1}$ represents the 5% and 10% trimmed means.
$\star$ indicates the best performing model in each column.

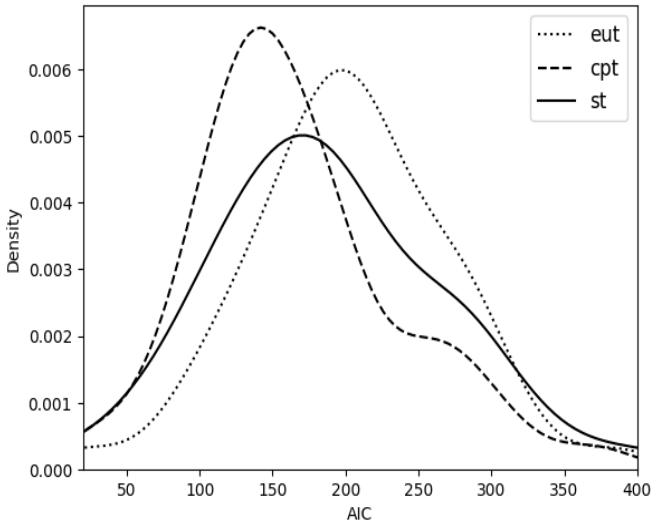
[17] Sugiura (1978) and Hurvich & Tsai (1989) modify the statistic for small-sample studies. The corrected AIC equals to $2k - 2ln(\mathcal{L}) + \frac{2k(k+1)}{n-k-1}$ where n is the number observations, and it punishes the model if there are too many parameters comparing to the sample size. We decide not to implement this correction since for each subject we have $8 \times 36 = 288$ observations, and the degrees of freedom are 1 (for EUT) and 2 (for CPT and ST).

[18] According to Burnham & Anderson (2002), Akaike weights are determined as follows. One wants to select the best model from N candidates (Model 1 to Model N), using Model b as the benchmark model (in practice, Model b is the one which is presumed to be the best). Firstly, one computes $\Delta_i = AIC_i - AIC_b$ for $i \in \{1, 2, 3, ..., N\}$, and AIC_b is the AIC value for model b. Then the Akaike weight $= \frac{\exp(-\Delta_i/2)}{\sum_{n=1}^{N} \exp(-\Delta_n/2)}$. The interpretation is straightforward: Akaike weights indicate the probability that a model is the best among the whole set of candidates.

Table 2.4: Ranking based on AIC

Theory	1st	2nd	3rd
EUT	18.75%	22.92%	58.33%
CPT	56.25%	37.50%	6.25%
ST	25.00%	39.58%	35.42%

Figure 2.2: Kernel density estimation of AIC



2.4.3 Predictive Power

Combined with the estimated model, the observations of each subject in Stage 2 are put back into Equation 2.13 to calculate the predicted log-likelihood which is used to measure the predictive power. The predicted log-likelihoods are analysed without any modification related to different degrees of freedom of theories, since overfitted models have disadvantages when comparing out-of-sample log-likelihoods (Hey et al., 2010). Table 2.5 reports the major statistics about predicted log-likelihoods. A similar pattern to descriptive power is found: CPT dominates among the three theories, and ST outperforms EUT.[19] This time, however, CPT outperforms ST by a very insignificant margin (-25.41 vs -25.98). Similarly, the Kernel density estimation (Figure 2.3) shows that the predicted log-likelihoods of EUT are centralised between roughly -30 and -25, and the predicted log-likelihoods of CPT and ST are decentralised with similar distributions (fat tail on the right side). Further, in Table 2.6, we report the percentage of subjects for whom each theory is dominating. Specifically, CPT wins for the majority

[19]On the one hand, all models outperform a random choice mechanism (on average). The predicted log-likelihood of a 'coin-tosser' (an agent who flips a coin every time he chooses between two alternatives) making 40 decisions is $ln(0.5) \times 40 \approx -27.73$. On the other hand, behaviours vary from individual to individual. In fact, there are around 29.2% of subjects who behave more similarly to a 'coin-tosser', rather than to an EUT agent. The percentages for CPT and ST are 12.2% and 22.9%.

of the subjects (60.42%), while for roughly half of the subjects (56.25%), ST is the second best, and EUT is 'the worst performer' for 58.33% of the subjects.

Further, Figure 2.4 shows the scatter plots of predicted log-likelihood against AIC for the three theories, and it demonstrates a negative correlation between the two statistics, i.e., positive correlation between descriptive and predictive power. The implication that theories with better descriptive ability are associated with higher predictive ability, in line with previous experiments, such as Hey et al. (2010), Hey & Pace (2014) and Georgalos (2019).

Table 2.5: Summary of predicted log-likelihoods

Theory	Mean	$\text{Mean}_{0.05}$	$\text{Mean}_{0.1}$	Median	s.d.
EUT	-27.20	-27.34	-27.35	-27.40	1.51
CPT	-25.41*	-25.11*	-25.27*	-26.12*	5.82
ST	-25.98	-26.01	-26.09	-26.63	2.81

Note: $\text{Mean}_{0.05}$ and $\text{Mean}_{0.1}$ represents the 5% and 10% trimmed means.
$\star$ indicates the best performing model in each column.

Table 2.6: Ranking based on predicted log-likelihoods

Theory	1st	2nd	3rd
EUT	18.75%	22.92%	58.33%
CPT	60.42%	20.83%	18.75%
ST	20.83%	56.25%	22.92%

Figure 2.3: Kernel density estimation of predicted log-likelihoods

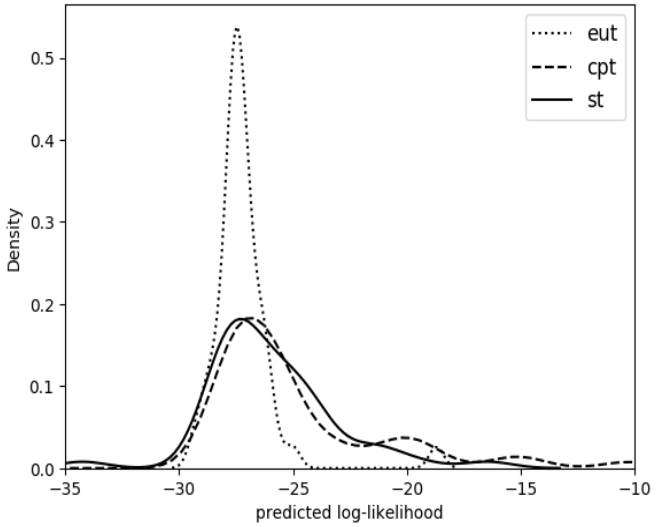
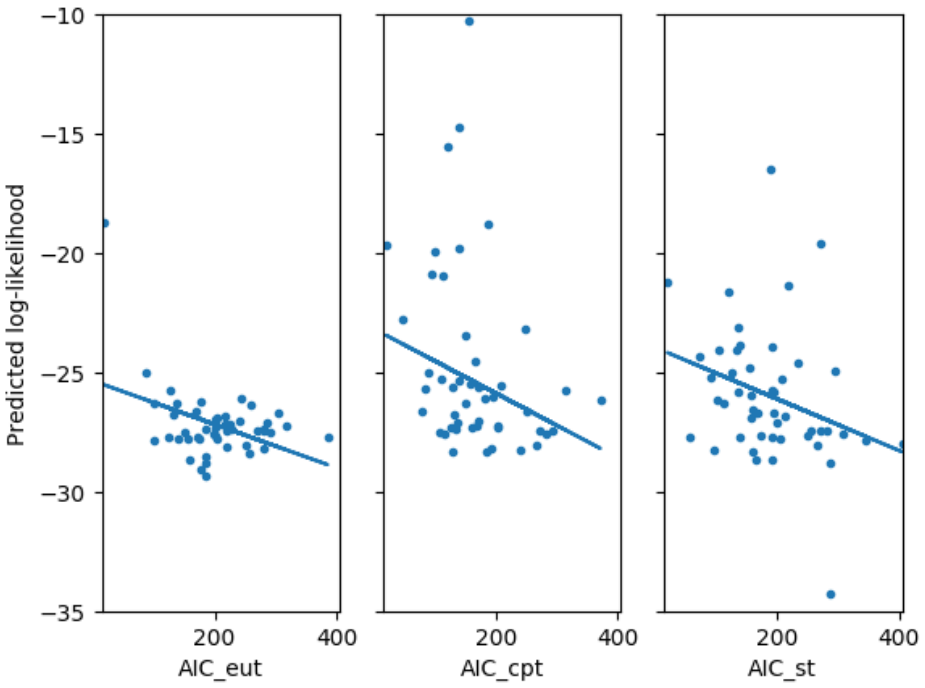


Figure 2.4: Scatter of predicted log-likelihood against AIC



2.4.4 Additional Analysis

Linear Utility vs Non-linear Utility: How Does It Matter?

The above results show that ST achieves predictive power similar to CPT, and it does not require a non-linear value function. In the framework of ST, most shifts of risk preferences can be explained by the 'ordering' and the 'diminishing sensitivity' properties of the salience function, as well as the function's convexity (Bordalo et al., 2012, p. 1278 -1279).[20] and the shape of the value function does not play a crucial role. However, the curvature of the value function is important for CPT. We examine how linear/non-linear utility affect the performance of both theories by introducing one variation of each. One is CPT with a linear value function (henceforth denoted as LCPT), and the other one is ST with a non-linear value function (henceforth denoted as NST). The parametric representation of LCPT is identical to CPT, except that the value function is linear, i.e., $v(x) = x$. The parametric representation of NST is identical to ST, except that the value function is non-linear, i.e., $v(x) = x^{\tau'}$, where $\tau' > 0$. For LCPT, the weighting function is $\frac{p^{\gamma'}}{(p^{\gamma'}+(1-p)^{\gamma'})^{1/\gamma'}}$, where $0 < \gamma' < 1$. For NST, the salience function is: $\frac{|x_i^s - x_j^s| + \theta'}{|x_i^s| + |x_j^s| + e\theta'}$, where $\theta' > 0$, and the local thinking parameter is $\delta' \in (0, 1]$. The summary of the estimated parameters and the distributions of the estimates are presented in Appendix A.3.

We report the average level of AIC and predicted log-likelihoods of LCPT and NST in Table 2.7. Comparing the result in Table 2.7 with the results in Table 2.3

[20] According to Bordalo et al. (2012), a salience function is convex if diminishing sensitivity becomes weaker as the overall payoff gets higher.

and Table 2.5, we see that the statistics for LCPT and NST roughly lie between CPT and ST. The pattern is clear for AIC, but not for predicted log-likelihoods, as the gap between CPT and ST is small in the first place. We shall take a close look at the Kernel density estimations in Figure 2.5. Despite the fact that the differences are small, it is unambiguous that, comparing to CPT, the descriptive and predictive performance of LCPT is closer to ST and NST does not perform better than ST. Also, it is worth noticing that in terms of applications, the calibration for LCPT is stabler than ST, and ST needs one extra degree of freedom. We believe that further work should focus on the parametric form of the salience function, since assuming non-linear utility does not improve performance.

Table 2.7: Summary of AIC and predicted log-likelihoods (LCPT and NST)

		Mean	$\text{Mean}_{0.05}$	$\text{Mean}_{0.1}$	Median	s.d.
LCPT	AIC	178.65	177.93	176.66	168.82	68.87
	predicted	-25.62	-25.76	-25.85	-26.84	3.47
NST	AIC	176.07	175.20	174.85	163.57	72.10
	predicted	-25.54	-25.69	-25.80	-26.64	3.20

Note: $\text{Mean}_{0.05}$ and $\text{Mean}_{0.1}$ represents the 5% and 10% trimmed means.

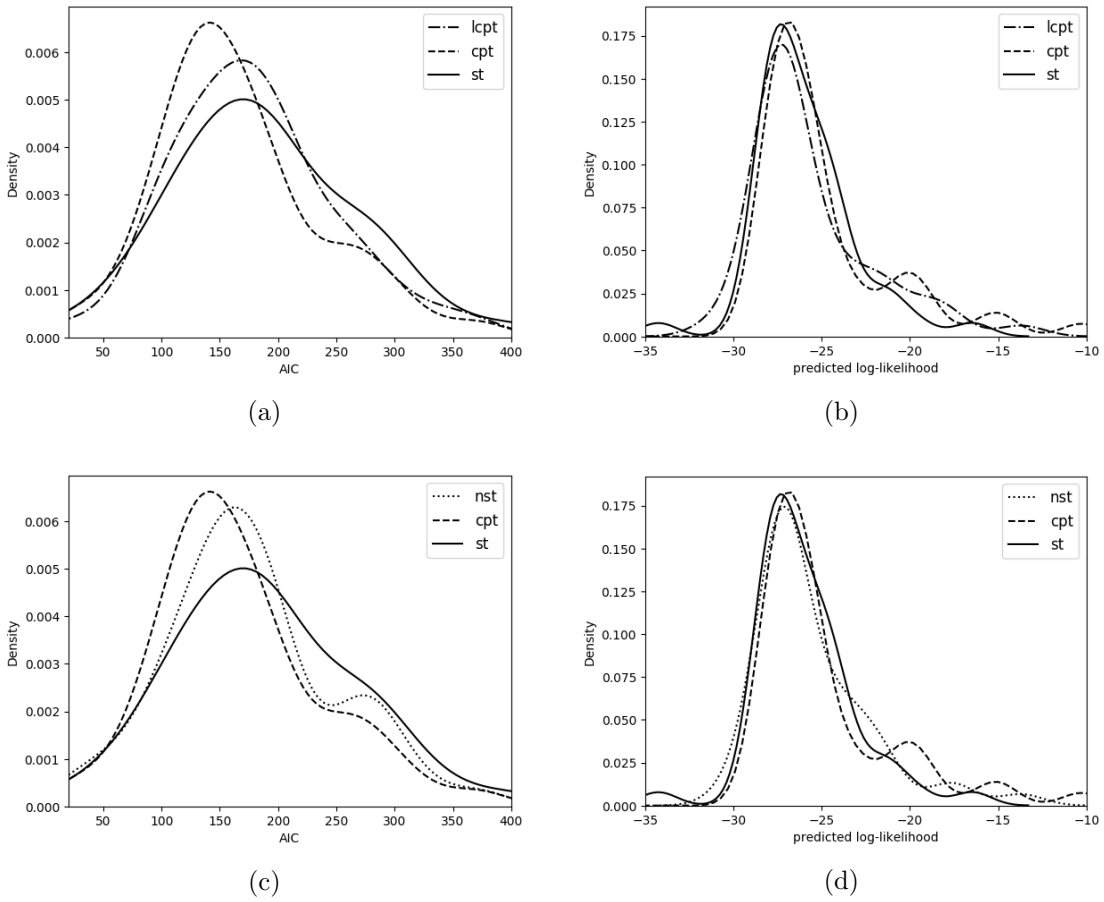
The Local Thinking Parameter

The local thinking parameter is δ in Equation 2.6, and it measures how much a local thinker differs from a rational economic decision maker (when $\delta = 1$, the decision maker does not distort the decision weights and ST reduces to expected value with the linear utility function). Königsheim et al. (2019) focus on this parameter and three of their main results are: i) the parameter is between 0.7 and 0.8; ii) the parameter does not change much when non-linear utility is assumed; iii) the parameter is not stable, in the sense that it is smaller if the choice question has a salient downside.[21] Our estimates of δ and δ' are about 0.7 which are in line with their first and second results (see Section 2.4.1 and Appendix A.3). However, further analysis reveals an exactly opposite result than their third result, i.e., we find heterogeneity, but a salient downside corresponds to significantly larger δ.

The design of Stage 1 provides a natural setting to examine this possible heterogeneity in δ. There are eight questions in each round of Stage 1, and four of them have a salient downside and the other four have a salient upside. Taking the sample question in Figure A.1 as an example, apparently, the upper four options in OPTION B column make the downside of OPTION A (receiving 0 pounds) more salient, while the bottom four options make the upside (receiving 20 pounds) more salient. We apply MLE to the observations of downside and upside questions separately, and the estimated

[21] A choice question is downside salient, if the low outcomes of the relative riskier lottery are in the most salient states. Therefore, when facing such question, a local thinker is more likely to focus on the downside of the riskier option and prefers the safer option.

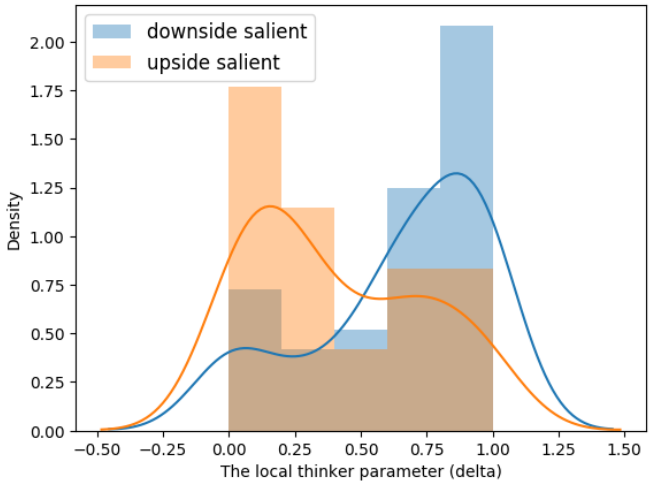
Figure 2.5: Kernel density estimations of AIC and predicted log-likelihood (LCPT and NST)



medians of local thinking parameters are 0.72 for questions with salient downside and 0.33 for questions with salient upside. In terms of statistical significance, the p-value of Wilcoxon rank-sum test is 0.0009, and the alternative is that values for the salient downside setting are more likely to be larger than the values in the upside setting. The distributions of the values in Figure 2.6 confirm this. We suggest that considering the heterogeneity of δ is a promising future direction for enhancing ST's predictive power.[22]

[22] A simple exercise using the observations in Stage 2 shows a slight increase in predictive power if we consider the heterogeneity of δ. Questions in Stage 2 can be categorised according to the 'directions of salience'. In terms of the 'mean-preserving' questions in Equation 2.11, $p \geq \frac{1}{2}$ indicates a salient downside and $p < \frac{1}{2}$ indicates a salient upside. In terms of the common consequence effect in Equation 2.12, a sufficiently small z indicates a salient downside (we consider z small if $z \in \{0, l/4\}$). Out of 40 questions in Stage 2, 25 of them are downside salient and the remaining questions are upside salient. If the predicted log-likelihoods of questions with different salience directions are calculated according to the corresponding calibrations, the mean of predicted log-likelihoods becomes -25.64 which is larger than the -25.98 of Table 2.5.

Figure 2.6: Distributions of the local thinking parameter: downside salient vs. upside salient



2.5 Conclusion

We empirically test Bordalo et al. (2012)'s ST in settings where the majority of subjects exhibit unstable risk attitudes. Firstly, we provide a formal calibration of ST, which has only been done by Königsheim et al. (2019). We find that the local thinking parameter is about 0.7, and the result is consistent with Königsheim et al. (2019)'s. Then, we compare ST and CPT on the basis of the in-sample and out-of-sample log-likelihoods. Two types of binary choice questions, which reveal the instability of subjects' risk preferences are used: 'mean-preserving' questions and common consequence effect questions. We find that CPT outperforms ST in terms of both descriptive and predictive power. However, the gaps are small, especially for predictive power. After a further investigation on the linearity of the value function and the heterogeneity of the local thinking parameter, we envision two possible pathways for improving ST for pure application purposes: developing a new specification of the salience function and analysing upside and downside salient settings with different local thinking parameters.

We shall conclude by an assessment of ST and its prospect as a new alternative to CPT. ST presents reasonable descriptive and predictive power, in comparison to the highly popular CPT. In our view, salience, in particular, the notion that people focus their attention on the most salient aspect of the world, plays a indisputable role in choice under risk. Salience can also be important for a variety of other economics situations, such as taxation (Chetty, Looney & Kroft, 2009), asset pricing (Bordalo, Gennaioli & Shleifer, 2013a), consumer behaviour (Bordalo, Gennaioli & Shleifer, 2013b) and judicial decisions (Bordalo, Gennaioli & Shleifer, 2015) etc. We anticipate that future empirical studies shall study salience in a broader domain of judgement setting.

Appendix A

A.1 Experimental Instruction

Introduction

Welcome and thank you for participating in today's experiment. It is important that you do not talk, or in any other way try to communicate during the study. You cannot use your phone during the study. If you have any questions, please just raise your hand and wait for the assistance.

During this experiment you will earn money. How much you earn depends on your decisions and luck. The money will be paid to you, in cash, at the end of the experiment.

Your participation in the experiment and any information about you will be kept anonymised and confidential. Your receipt of payment and consent form are the only places on which your name will appear. This information will be kept confidential in the manner described in the consent form.

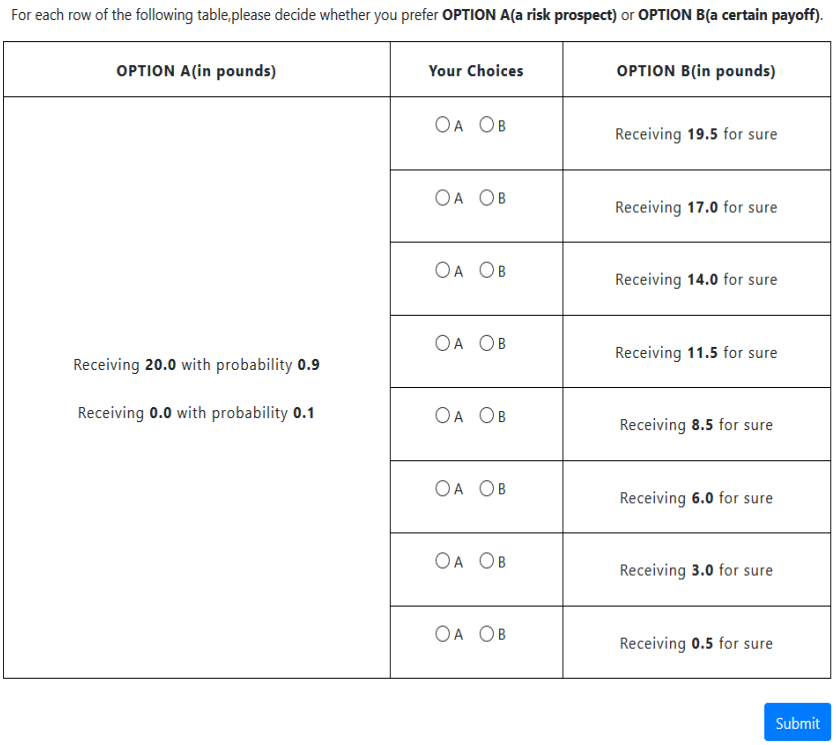
The experiment has two stages. Stage 1 consists of 36 periods. In each period, you will make choices between a lottery and a series of certain values. In Stage 2, you will make 30 choices. There will be a short questionnaire at the end.

Stage 1

The first stage will have 36 periods. In each period, the screen displays a lottery (Option A) on the left and displays a descending series of eight sure outcomes (Option B) on the right side. Every period has the same general form.

An example of the screen for a period is given in Figure A.1. As you can see, for each decision, you must choose between Option A and Option B. You may choose Option A for some decisions and Option B for others, and you may change your decisions and make them in any order. Once you have made all of your decisions, press the Submit button and you will be taken to the next period. Note that after you have pressed the submit button, you will no longer be able to change your decisions.

Figure A.1: Stage 1 sample screen

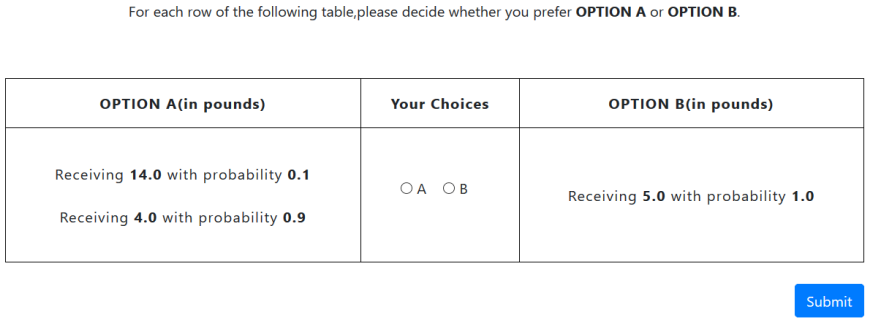


For each row of the following table,please decide whether you prefer **OPTION A(a risk prospect)** or **OPTION B(a certain payoff)**.

OPTION A(in pounds)	Your Choices	OPTION B(in pounds)
Receiving **20.0** with probability **0.9** Receiving **0.0** with probability **0.1**	○A ○B	Receiving **19.5** for sure
	○A ○B	Receiving **17.0** for sure
	○A ○B	Receiving **14.0** for sure
	○A ○B	Receiving **11.5** for sure
	○A ○B	Receiving **8.5** for sure
	○A ○B	Receiving **6.0** for sure
	○A ○B	Receiving **3.0** for sure
	○A ○B	Receiving **0.5** for sure

Submit

Figure A.2: Stage 2 sample screen



For each row of the following table,please decide whether you prefer **OPTION A** or **OPTION B**.

OPTION A(in pounds)	Your Choices	OPTION B(in pounds)
Receiving **14.0** with probability **0.1** Receiving **4.0** with probability **0.9**	○A ○B	Receiving **5.0** with probability **1.0**

Submit

Stage 2

In this stage, you will need to make 30 choices. An example of the screen for a question is given in Figure A.2. As you can see, you simply choose between Option A and Option B, and press the Submit button to go to the next question. Similar to Stage 1, you cannot change your decision after you have pressed the submit button.

Your payment

- You will receive a participation fee of £4 regardless of your decisions.

- You can get an additional Stage 1 fee equal to £4. Note that, if you complete Stage 1, however leave the experiment before Stage 2 is finished, you will only earn the participation fee of £4.
- After you complete Stage 2, the computer will randomly choose a question from Stage 2 and you will be paid according to this question. Depending on the random question and the decision of yours, there are two possibilities:
  1 If the Option you chose in the random question is paying an amount for sure, then that amount will be your payment in Stage 2.
  2 If the Option you chose in the random question is a lottery, you will roll a ten-sided die to determine your payoff. For example, suppose that that Option is as follows:

$$\begin{cases} \text{Receiving £25 with probability 0.1} \\ \text{Receiving £5 with probability 0.9} \end{cases}$$

  In this case, you will roll the ten-sided die once. If a 1 comes up, then you will receive £25 , while if a 2, 3, 4, 5, 6, 7, 8 ,9 or 0, comes up then you will receive £5.
  As a further example, suppose that the probabilities in that Option have two decimal digits:

$$\begin{cases} \text{Receiving £20 with probability 0.33} \\ \text{Receiving £15 with probability 0.66} \\ \text{Receiving £0 with probability 0.01} \end{cases}$$

  In this case, you will roll the die twice. These two rolls will correspond to a number from 00 to 99. For instance, if you roll 0, 4, it corresponds to 04; if you roll 4, 0, it corresponds to 40. In this example, if the corresponding number you roll is 01, 02, . . . , or 33, then you will receive £20. If your roll is 34, 35, . . . ,or 99, then you will receive £15. If you roll 00, then you will receive £0.
- Your payment = Participant fee (£4) + Payment in Stage 1 (£4) + Payment in Stage 2 (£0 - £15)

If you have any questions, please raise your hand now, otherwise we will begin with the experiment.

A.2 Questions in Stage 2

§ indicates 'mean-preserving' questions. † indicates upside salient questions. Note that in real sessions, we randomised the display order of the questions for each subject. However, the left-right juxtaposition remained the same as a consequence of software

limitations.

1. $L = (6, 1), \quad R = (12, 0.4; 2, 0.6)$§†
2. $L = (6, 0.5; 4; 0.5), \quad R = (5, 1)$§
3. $L = (7, 1), \quad R = (8.5, 0.4; 6, 0.6)$§†
4. $L = (7, 1), \quad R = (9, 0.6; 4, 0.4)$§
5. $L = (7, 1), \quad R = (12, 0.5; 2, 0.5)$§
6. $L = (8, 1), \quad R = (12.5, 0.4; 5, 0.6)$§†
7. $L = (8, 0.5; 6, 0.5), \quad R = (7, 1)$§
8. $L = (9, 1), \quad R = (15, 0.25; 7, 0.75)$§†
9. $L = (9, 0.25; 5, 0.75), \quad R = (6, 1)$§†
10. $L = (9, 0.5; 5, 0.5), \quad R = (7, 1)$§
11. $L = (9, 0.76; 6.5, 0.24), \quad R = (10, 0.75; 6.5, 0.24; 0, 0.01)$
12. $L = (10, 1), \quad R = (11, 0.8; 6, 0.2)$§
13. $L = (10, 1), \quad R = (11.5, 0.4; 9, 0.6)$§†
14. $L = (10, 0.11; 2.5, 0.89), \quad R = (12, 0.1; 2.5, 0.89; 0, 0.01)$†
15. $L = (10, 0.55; 2.5, 0.45), \quad R = (12, 0.5; 2.5, 0.45; 0, 0.05)$†
16. $L = (10, 0.11; 5, 0.89), \quad R = (11, 0.1; 5, 0.89; 0, 0.01)$
17. $L = (10, 0.76; 5, 0.24), \quad R = (11, 0.75; 5, 0.24; 0, 0.01)$
18. $L = (10, 0.75; 6, 0.2; 0, 0.05), \quad R = (8, 0.8; 6, 0.2)$
19. $L = (10, 0.5; 6.5, 0.49; 0, 0.01), \quad R = (9, 0.51; 6.5, 0.49)$
20. $L = (10, 0.34; 7.5, 0.66), \quad R = (11, 0.33; 7.5, 0.66; 0, 0.01)$
21. $L = (11, 1), \quad R = (12, 0.5; 10, 0.5)$§
22. $L = (12, 0.11; 0, 0.89), \quad R = (13, 0.1; 0, 0.9)$†
23. $L = (12, 1), \quad R = (15, 0.8; 0, 0.2)$§
24. $L = (12, 0.4; 7, 0.6), \quad R = (9, 1)$§
25. $L = (12, 0.8; 7, 0.2), \quad R = (11, 1)$§
26. $L = (13, 0.9; 0, 0.1), \quad R = (12, 0.95; 0, 0.05)$†
27. $L = (13, 0.5; 2.5, 0.45; 0, 0.05), \quad R = (11, 0.55; 2.5, 0.45)$†
28. $L = (13, 0.11; 3, 0.89), \quad R = (15, 0.1; 3, 0.89; 0, 0.01)$†
29. $L = (13, 0.11; 6.5, 0.89), \quad R = (15, 0.1; 6.5, 0.89; 0, 0.01)$
30. $L = (13, 0.33; 8, 0.66; 0, 0.01), \quad R = (11, 0.34; 8, 0.66)$
31. $L = (13, 0.51; 9.5, 0.49), \quad R = (14, 0.5; 9.5, 0.49; 0, 0.01)$
32. $L = (14, 0.9; 0, 0.1), \quad R = (13, 0.95; 0, 0.05)$†
33. $L = (14, 0.75; 6.5, 0.24; 0, 0.01), \quad R = (13, 0.76; 6.5, 0.24)$
34. $L = (14, 0.33; 9.5, 0.66; 0, 0.01), \quad R = (13, 0.34; 9.5, 0.66)$

35. $L = (14, 0.75; 9.5, 0.24; 0, 0.01), \quad R = (13, 0.76; 9.5, 0.24)$
36. $L = (14.5, 0.8; 2, 0.2), \quad R = (12, 1)$§
37. $L = (15, 0.9; 3.5, 0.09; 0, 0.01), \quad R = (14, 0.91; 3.5, 0.09)$†
38. $L = (15, 0.1; 5, 0.9), \quad R = (6, 1)$§†
39. $L = (15, 0.25; 11, 0.75), \quad R = (12, 1)$§
40. $L = (15, 0.5; 13, 0.5), \quad R = (14, 1)$§

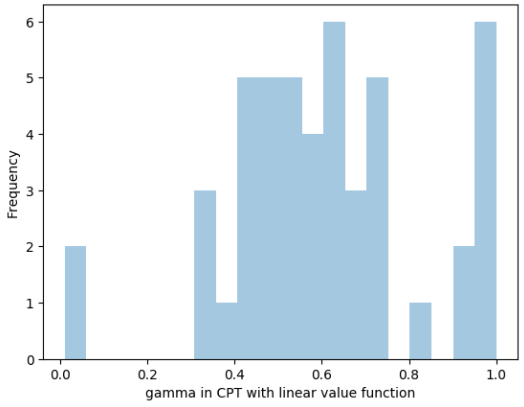
A.3 Additional Tables and Graphs

Table A.1: Descriptive Summary of Estimates (LCPT and NST)

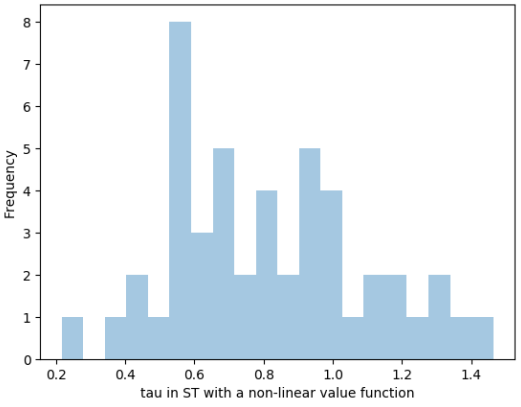
Theory	Parameter	Mean	Median	s.d.
LCPT	γ'	0.60	0.58	0.22
	s_{lcpt}	4.53	3.32	3.94
NST	τ'	0.82	0.81	0.28
	θ'	72.94	1.97	202.62
	δ'	0.66	0.77	0.32
	s_{nst}	3.98	2.26	4.60

Note: The results exclude one subject for the reasons explained in footnote 14.

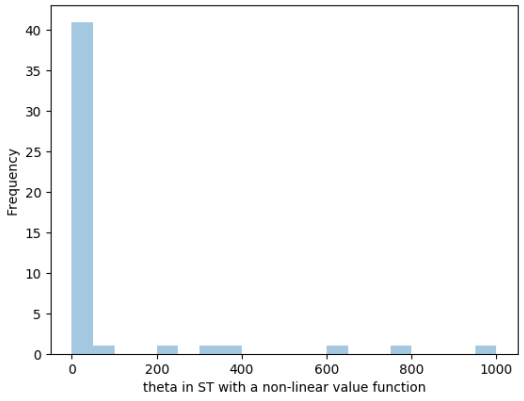
Figure A.3: Distributions of estimates for LCPT and NST



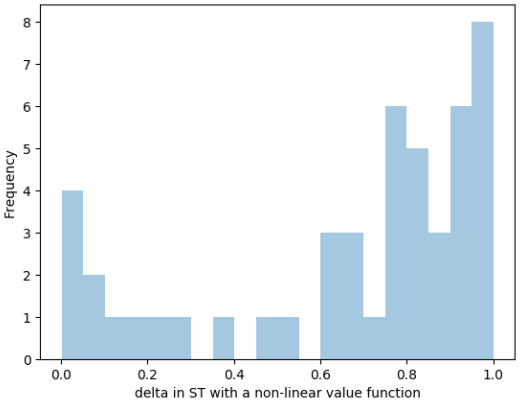
(a)



(b)



(c)



(d)

Chapter 3

How Economically Important is Anchoring? A Research Synthesis of WTP/WTA Studies

3.1 Introduction

Anchoring can be generally defined as the influence of a normatively irrelevant cue on a subsequent expression of judgement. In particular, the cue and the judgement can be numerical. Since the pioneering work of Tversky & Kahneman (1974) (henceforth denoted as TK), numerical anchoring has been considered one of the most robust psychological phenomena in judgement and decision-making. In the seminal study of TK a roulette wheel was used to deliver a random number, on the basis of which subjects were asked a binary question about some unknown quantity. For instance, "does the average temperature in Antarctica exceed the random number drawn from the roulette wheel?". Subsequently, subjects were requested to make judgments about the actual magnitude of the given variable (in this case, the average temperature in Antarctica). TK found that the elicited numeric judgements were greatly affected by the initial binary question. In particular, if a subject drew a large random number (in the support of possible numbers) they tended also to express higher magnitudes in their numerical judgement tasks.

Anchoring belongs to the domain of behavioural research termed 'heuristics and biases' by TK, in which consumers tend to deviate systematically from the benchmark of rational economic behaviour. A key question is the economic importance of anchoring. A fundamental postulate of economic theory is the concept of 'consumer preferences' that shape economic behaviour and are the foundation of demand and thus market prices. Demand is the expression of willingness-to-pay (WTP) for economic goods. Importantly, anchoring has been shown to matter in the elicitation of economic preferences, with early demonstrations in the works of Northcraft & Neale (1987), Green et al. (1998) and Ariely et al. (2003). Ariely et al. (2003) employed the prototypical design of TK to elicit the

WTP for a series of consumer goods as well as the Willingness-to-Accept (WTA) for negative simple hedonic experiences.[1]

In particular, the prototypical design of anchoring in economic evaluation consists of three main stages. The first stage is what we call an *anchoring manipulation*, where the experimenter shows subjects a good and asks: "Would you purchase (sell) this object for 'X' Dollars?" where 'X' is the numerical anchor. The second stage is the *elicitation* of the subject's economic valuation. For instance, (for eliciting WTP) the experimenter might ask the subject a question such as "What is the maximum dollar amount would you be willing to pay for this object?" The last stage provides incentives for truthfully revealing the valuation, (for instance, using the Becker-Degroot-Marschak Mechanism that emulates a real market). A large number of subsequent experimental studies followed this prototypical design and most of them have the first and second stage (which might take different formats). Not all studies have the incentivisation stage.

As a consequence of these early results, not only is anchoring considered relevant for economic preferences, but also extremely robust across several dimensions, and its magnitude is believed to be large. However - perhaps because of the robustness of the phenomenon in the psychology literature - the exact economic magnitude of anchoring has not been given enough focus. Most subsequent studies found weaker and less robust anchoring than early influential studies, but the totality of the literature has not been synthesised into a statement about the quantitative economic significance of numerical anchoring.

There are important benefits from encapsulating what is known and quantifying the effect of anchors on consumers' economic valuation. First, an assessment of whether the effect is large enough to be a concern would be useful for methodological appraisals of the contingent valuation methodology as well as for public policy aiming to protect consumers. Marketers would also like to know how malleable their customers' WTP is. Of course, information about factors that moderate this influence would be equally important. Some key questions from the economic point of view are: What types of goods are most prone to anchoring? Do market forces ameliorate anchoring? From the methodological point of view, economists would also like to know whether certain methodologies (such as monetary incentives) tend to be associated with higher or lower anchoring effects.

Examining the factors which affect the magnitude of anchoring may illuminate the causes of anchoring as well. The most well-known theory of anchoring is TK's 'anchoring and adjustment' which states that people consider the arbitrary cue as a possible answer to the evaluation question. They treat the anchor as a starting point and then adjust, but the adjustment is always insufficient. Another popular model is the 'selective accessibility' model, and Mussweiler, Strack & Pfeiffer (2000) summarise that the anchoring manipulation increases the accessibility of anchor-consistent knowledge which is used for the later evaluation. Although it's not our main focus, the results

[1] For instance, such experiences entail hearing annoying but harmless sounds through headphones, or drinking bad-tasting but harmless liquids.

of the moderator analyses shall have implications regarding the theoretical aspects of anchoring.

This study aims to fill these gaps by performing a systematic synthesis of studies that examine the effects of numerical anchors on statements of economic valuation. Our task is to systematise the measurements of anchoring quantitatively, as well as to study the determinants of variability in the magnitude of the effect size. We include 53 studies from 24 articles and choose the Pearson correlation coefficient between the anchor number and target response (in our case, WTP/WTA) as the primary effect size. Both fixed-effects and random-effects models point to a moderate overall effect. Further meta-regression analysis shows that subjects in WTP tasks are more likely to be influenced, comparing to subjects in WTA. Incentives do not attenuate the effects. Also, the relevance and compatibility of the anchor to the target response and the location of the experiment matter for the magnitude of the effect size. Overall, the effect of anchoring on economic evaluation should not be overlooked, but it does not seem to be as strong as previously believed.

The remainder of this document is as follows. Section 3.2 discusses our design and methodological choices, in particular our search of the literature, as well as the choice of 'effect size' measure and of moderator variables. Section 3.3 reports the standard meta-analytic results which include the overall effect size and the meta-regression results. Section 3.4 examines the robustness of our result. Section 3.5 concludes.

3.2 Methods

3.2.1 Effect Sizes

A key methodological choice of our study is the main effect size measure. The effect size that we analyse is the Pearson correlation coefficient (r) between anchor number and target response (in our case WTP/WTA). We have chosen this effect size because its interpretation is natural in our economic setting, it is reported in several studies in the included literature and it is a standard meta-analytic measure (Cooper, Hedges & Valentine, 2009).[2]

For studies where the raw data is not available, we extract the effect size using reported information. We employ standard meta-analytic methodology for translating reported effect size measures into alternative ones (Borenstein, Cooper, Hedges & Valentine, 2009), following the guidelines in Cooper et al. (2009) (Chapter 12, Page 224 - 234)

[2]We considered two other possible measures: Spearman's rank correlation coefficient and Jacowitz & Kahneman (1995)'s 'anchoring index'. The latter can be defined as follows: For a binary between-subject treatment, where one sample of subjects has been exposed to a low anchor value and another sample to a high anchor value: AI= [Median (High Anchor)- Median (Low Anchor)]/[High Anchor- Low Anchor]. Both of these measures have desirable properties in terms of measuring the magnitude of the anchoring effect. Spearman's rank correlation coefficient captures monotonicity in cases where the effect is not linear and anchoring index offers an intuitive descriptive metric. However, we need raw data to calculate any of these measures, and raw data is available only for less than half of the included studies (24/53).

to transform the given measure into r. The Campbell Collaboration online effect size calculator Wilson (2001) is also used as complementary tool.[3]

At this point we need to make an important methodological aside. Many studies report the elicitation of valuations for multiple goods. Several designs are within-subject, which implies that often a given subject is faced with multiple anchors. Since the psychological processing of multiple anchors is complex and different from the effect of a single anchor (Whyte & Sebenius, 1997), we wish to focus on the effect of single anchors. This requires some attention to isolate the data that originate from exposure to single vs. multiple anchors. This may not always be possible given the information contained in each article. We shall get back to this important point.

3.2.2 Literature Search and Inclusion Criteria

We retrieved the studies using the following four channels. 1) We searched the Web of Science, Google Scholar and EconLit databases.[4] 2) We used Web of Science to focus on the references and the citations of the aforementioned three early studies (Northcraft & Neale, 1987; Green et al., 1998; Ariely et al., 2003). 3) Personal communication with specialist researchers also helped to trace unpublished articles. 4) We posted a literature searching advertisement on the ESA Experimental Methods Discussion group.[5]

After the studies were retrieved, we screened the papers, and studies are included by mutual agreement on the basis of prespecified inclusion criteria as follows. Only English-speaking studies were considered. We looked for studies eliciting a numerical statement of WTP/WTA for economic goods after an unambiguous and unique numeric anchor is presented. Some examples of inclusions and exclusions are provided as follows. 'A numerical statement of WTP/WTA' excludes studies like Wansink, Kent & Hoch (1998) who only measure quantities purchased, Jung, Perfecto & Nelson (2016) who elicit 'pay what you want' which in our view is not a representation of WTP, Mussweiler et al. (2000) who asked subjects to perform a neutral pricing task,[6] etc. We consider studies which present 'an unambiguous and unique numeric anchor'. Therefore, we exclude studies with explicit multiple anchors, such as Sugden, Zheng & Zizzo (2013) where a within-subject design was used, in which a given subject is exposed to different goods and different anchors in a random order. We also exclude studies that used the

[3] For most studies for which raw data are not available, we can derive the effect size using the aforementioned methods. The only exceptions are studies which only report the coefficient of a multiple regression. For those studies, we used the formula provided in Peterson & Brown (2005). The formula is $r = 0.98\beta + 0.5\lambda$, where $\lambda = 1$ if β is non-negative, otherwise $\lambda = 0$.

[4] The following search string was used in Web of Science: (TS=(anchoring AND willingness to pay) OR TS=(anchoring AND willingness to accept) OR TS=(anchoring AND valuation) OR TS=(anchoring AND "WTA") OR TS=(anchoring AND "WTP")) AND LANGUAGE: (English). In EconLit, we searched for similar keywords.

[5] The ESA Experimental Methods Discussion group is a Google group for economists to discuss experimental methods in economics, and it is sponsored by the Economic Science Association.

[6] In their experiment, the following question is asked in the elicitation phase: "Could you tell me, what do you think is the approximate price for the car as you see it?". Our interpretation is that this elicitation is not of WTP or WTA.

'list method' to elicit WTA/WTP, such as Araña & León (2008) and Tufano (2010).[7] In general, if in a study it is possible to identify the first anchor to which a given subject was exposed, we include the given study and calculate the effect size using the first anchor and the corresponding elicited WTA or WTP. Therefore, we include studies such as Bavolár (2017), where subjects are presented with different anchors and goods but it is easy to identify that the first good is a 'pasameter' and Green et al. (1998), where subjects are presented with five numerical evaluation questions with the first one being a WTP evaluation.[8] Our included studies consider not only market goods, but also lotteries (Fudenberg et al., 2012), environmental goods (Green et al., 1998; Schläpfer & Schmitt, 2007), simple hedonic experiences (Maniadis et al., 2014), etc.

3.2.3 Moderators

Our methodological objective is to code for theoretically relevant aspects of the design that have an influence on the estimated effect sizes. Except for the standard 'moderator variables' in a meta-analytic study, such as the sample size, the year of publication, the country of study and the subject pool (students vs. general population), we coded the following seven moderators: anchor type, good type, task type, incentive type, experiment type, compatibility, and manipulation type (see Table 3.1 for a summary of the moderators and categories).

Anchor Type

Ariely et al. (2003) and many of its replications used explicitly random anchors (for instance, the last two digits of a subject's social security number). Also, there are experiments that provided a fixed anchor number without an explanation of its origin. In several studies, subjects are given anchors which have some potential relevance with the target. For instance, in Bavolár (2017) the anchor is provided as a price paid by a hypothetical person. Here our theoretical prediction is that higher relevance will be associated with a stronger anchoring effect. The reason is that the anchor may convey true information about the underlying properties of the good, and in the case of a fixed anchor, subjects might assume that the anchor number is provided for a reason.

Task Type

Another variable is the type of elicitation task (WTA vs. WTP). There are important differences in the way that people express their WTP and WTA, and there is a famous 'gap' between the two (Kahneman, Knetsch & Thaler, 1991). In pure WTP and WTA tasks, WTA is usually found larger than WTP. However, the literature on the

[7] The 'list method' asks subjects repeatedly (in the elicitation phase) whether they would buy or sell an object for different prices. These prices are all salient at the time of final choice and therefore can themselves serve as anchors. We therefore chose to exclude these studies.

[8] Note that in every case, we include an effect size wherever there is an elicitation of valuation after exposure to a single anchor, and we can unambiguously determine the anchor and the corresponding elicitation.

Table 3.1: Summary of Moderators

Moderators	Categories	Coding
Anchor Type (AT)	Explicitly random	0
	Fixed and provided without explanation	1
	Having some relevance with the target	2
Task Type (TT)	WTP	0
	WTA	1
Good Type (GT)	Easy to evaluate	0
	Difficult to evaluate	1
Manipulation Type (MT)	Canonical design	0
	Non-Canonical design	1
Subject Pool (SP)	Students	0
	General population	1
Incentives Type (IT)	Not incentivised	0
	Probabilistically incentivised	1
	Fully incentivised	2
Experiment Type (ET)	Lab experiment	0
	Class experiment	1
	Field experiment	2
Compatibility (CP)	Compatible	0
	Incompatible	1

possible disparity of anchoring effect across these two tasks is insufficient. Simonson & Drolet (2004)'s experiments suggest that WTP is more susceptible to anchoring effect compared to WTA. They argue that this is because in WTA tasks sellers set prices based on the market price, which is objective, while buyers' subjective values of the goods play an important role in WTP tasks. If Simonson & Drolet (2004)'s argument holds, we should expect the effect sizes for WTA studies to be smaller than those eliciting WTP.

Good Type

The literature has considered several types of economic 'goods' for which WTP/WTA is elicited. In general, we should expect that some types of goods would be associated with a stronger capacity to retrieve one's underlying preferences, and hence lower tendency to be affected by anchoring (Ariely et al., 2003). The difficult task is to ascertain the types of goods used in each particular study. After careful consideration, we decided to split the goods into two general groups: 'easy goods' which are in general easier for one to construct their WTA or WTP, vs. 'difficult goods'. The former category includes normal market goods such as chocolates and wine, as well as simple hedonic experiences and lotteries.[9] The latter contains objects such as luxury hotels (Tanford, Choi & Joe, 2019), environmental goods (Green et al., 1998; Schläpfer & Schmitt, 2007), or real estate (Northcraft & Neale, 1987), etc. These goods are complex items, which the average experimental participant should not be expected to use more than once in three years. Of course, our hypothesis is that goods of the first category will be easier for subjects to evaluate and hence will be less prone to anchoring effects.

Incentive Type

Another important variable is the magnitude of monetary incentives. Using financial incentives is considered a methodological norm in economic experiments and according to this perspective we should expect that they incentivise accurate statements of economic valuation. However, there is much heterogeneity in the literature regarding this design aspect, and the majority of studies are not incentivised - especially for experiments which involve expensive goods. Many studies only picked randomly a small number of participants in a given experimental session for whom one of the choices is consequential (probabilistically incentivised). Only a few studies have at least one incentivised decision for each participant (fully incentivised).

Experiment Type

According to the physical location of the experiment, we categorise three types of experiments: lab experiment, class experiment and field experiment. The hypothesis here is that different types of experiments are associated with different levels of experimental control. In particular, a lab is a more controlled environment than a class, which is more controlled than the field. If we find that this variable impacts the effect size, then we might say something about the robustness of the phenomenon.

[9]The frequent consumption of normal market goods should clearly facilitate the numerical valuation of these goods. We consider also the last two categories of goods 'easy' for the following reasons. First, people have experience with lotteries and the ones used in the lab usually have an expected value easy to calculate. Secondly, as Ariely et al. (2003) argue, simple hedonic experiences provide direct access to the 'pleasures and pains' associated with consumption, so it should be easy to evaluate the monetary equivalent of them.

Compatibility

We also coded for a variable which – as the psychology literature teaches us – might be relevant for anchoring. The variable is called 'compatibility' between the anchor and the target, indicating whether the anchor and evaluation are expressed in the same dimension and the same units (Strack & Mussweiler, 1997). WTA and WTP are in money units, thus we coded the study as "compatible" as long as the anchor is expressed in the monetary dimension and in the same units as the elicitation. [10]

Manipulation Type

The manner in which the anchoring manipulation is operationalised is also important. In particular, we coded experiments which do not use the *anchoring manipulation* stage followed by the *elicitation* stage (the aforementioned prototypical design described in Section 3.1) as having a 'non-canonical' design.[11]

We experimented with several other potentially relevant variables. These include whether the anchor is plausible (Mussweiler & Strack, 2001), the elicitation method of WTP/WTA (an 'open-ended' question or some form of auction), measures of emotions (Araña & León, 2008), forewarnings about the role of anchoring (Epley & Gilovich, 2005; LeBoeuf & Shafir, 2009), etc. We chose not to include these variables because in the process of coding we realised that there is not enough heterogeneity.[12]

A word of caution related to this choice. Thompson & Higgins (2002) emphasise that the results of meta-regression analysis will necessarily be correlational, not causal, and warn against 'data dredging', namely, examining multiple models and post-hoc theorising. This is particularly problematic in meta-analysis because it uses the totality of the evidence and thus is not possible to validate a model with out-of-sample predictions. In order to discipline ourselves, we drop variables that do not show sufficient heterogeneity, and we shall hypothesise carefully about the potential mechanisms, keeping a small number of key moderators before embarking in our meta-regression.

[10] Only a few studies, such as Dogerlioglu-Demir & Koçaş (2015), Schläpfer & Schmitt (2007) and Tanford et al. (2019) used incompatible anchors. In particular, Dogerlioglu-Demir & Koçaş (2015) used as anchors some numbers that appear in the name of a given good (for instance, subjects evaluate an average meal at 'Studio 17 versus Studio 97'); Schläpfer & Schmitt (2007) use tax rates presented in the form of percentages as anchors; Tanford et al. (2019) claim that one of their treatments is incompatible, since the economic evaluation is the price for a hotel 'per night', and the anchors are presented in 'per week' terms.

[11] For instance, Yu, Gao, Sims & Guan (2017) simply put a label with the anchor number on the goods and did not ask the comparative question; Northcraft & Neale (1987) and Tanford et al. (2019) gave to participants the listing price for the goods; Bavolár (2017) introduce a stage where they present the anchor number, but not in the form of a question, etc.

[12] For all included studies, no information was given regarding anchor plausibility, emotions or forewarnings, and only two studies - one in Ariely et al. (2003) and one in Yu et al. (2017) - used auctions to elicit WTP/WTA.

3.3 Meta-Analytic Results

3.3.1 Description of Studies

We included 24 articles, and since several articles contain multiple studies, there are in total 53 studies.[13] We include articles in which the author(s) conducted several experiments, and articles which contain one experiment with multiple conditions (moderators). In the latter case, we treat the article as containing several studies based on those conditions. Some detailed information is reported in Table 3.2. We treat studies contained in a single article as independent. In the practice of meta-analysis, this methodological choice is reasonable, so as long as the studies use different subjects, and there is no issue of counting subjects multiple times. We assume that mere 'article-level effects' or 'author-level effects' (bias stemming from the fact that different studies are conducted by overlapping sets of authors) are insignificant. However, we shall provide relevant robustness tests after the main analysis.

We are also able to get raw data for 13 out of 24 articles (24 out of 53 studies).[14] For Ariely et al. (2003) and Simonson & Drolet (2004), only partial data is available.[15] Figure 3.1 provides the general picture: it shows the numbers of studies falling into each category of the different moderators. In general, all moderators show heterogeneity to a certain extent. Figure 3.2a shows that there was limited work on the anchoring effect on economic evaluation up until the late 90s, and after Ariely et al. (2003)'s work the topic became particularly popular. Figure 3.2b illustrates the distribution of sample sizes in the literature and shows that most studies had a sample size smaller than 200.

[13] In terms of the use of the term 'study' in this paper, it represents the 'unit' included in this meta-analysis. Therefore, we say that we have 53 studies, and a study can be an experiment or a condition in one experiment.

[14] We requested the data by sending emails to the researchers and no reminders were sent.

[15] For Ariely et al. (2003), we have the data for "EXPERIMENT 1: COHERENTLY ARBITRARY VALUATION OF ORDINARY PRODUCTS" from page 75. For Simonson & Drolet (2004), we have data for 'Study 1' from page 683.

Table 3.2: Summary of included articles

Article	# of Studies	Method	Raw data
Adaval & Wyer (2011)	2	2	Yes
Alevy, Landry & List (2015)	2	2	Yes
Ariely et al. (2003)	4	1	Partial
Andrersson & Wisaeus (2013)	1	NA	No
Bavolár (2017)	1	NA	Yes
Bergman, Ellingsen, Johannesson & Svensson (2010)	1	NA	Yes
Brzozowicz, Krawczyk, Kusztelak & others (2017)	2	2	Yes
Brzozowicz & Krawczyk (2019)	2	2	Yes
Dogerlioglu-Demir & Koçaş (2015)	4	2	No
Fudenberg et al. (2012)	4	1	Yes
Green et al. (1998)	1	NA	No
Koçaş & Dogerlioglu-Demir (2014)	1	1	No
Li, Fooks, Messer & Ferraro (2019)	1	NA	No
Maniadis et al. (2014)	1	NA	Yes
Northcraft & Neale (1987)	4	1 and 2	No
Nunes & Boatwright (2004)	1	NA	No
Schläpfer & Schmitt (2007)	1	NA	Yes
Simonson & Drolet (2004)	8	1, 2	Partial
Tanford et al. (2019)	2	1 and 2	Yes
Wu, Cheng & Lin (2008)	2	1	No
Wu & Cheng (2011)	2	2	No
Yoon, Fong & Dimoka (2013)	1	NA	Yes
Yu et al. (2017)	1	NA	No
Yoon & Fong (2019)	4	1	No

The Method column - 1: one article breaks into several studies because of multiple experiments conducted; 2: one experiment breaks into several studies on the basis of different moderators.

3.3.2 Average Effect Size

In Figure 3.3 we present a general overview of the extracted effect sizes (correlation coefficient). As we can see from the figure, there is a negative relationship between the publication year and the magnitude of the effect size. In the second panel we plot the effect size against the study sample size. This illustrates a weak positive relationship.[16]

In practice, meta-researchers usually do not perform the meta-analysis directly with r, but r is transformed to Fisher's z using this formula:

$$z = 0.5 \times ln(\frac{1+r}{1-r}) \tag{3.1}$$

[16]It is worth noting that there is one study which has a particularly high effect size (larger than 0.8). This is the work of Green et al. (1998), who conducted a large field experiment to elicit the WTP of an environmental good. Since this is a particularly extreme outlier, we also consider the robustness of the results to excluding this outlier, especially in the meta-regression part. In particular, in Appendix B.3 we provide meta-analytic results where this study is excluded. Comparing to the results of the complete dataset, the overall effect size is smaller but the significance levels of the coefficients in the regression are very similar.

Figure 3.1: Summary of number of studies in each category, complete dataset (53 studies)

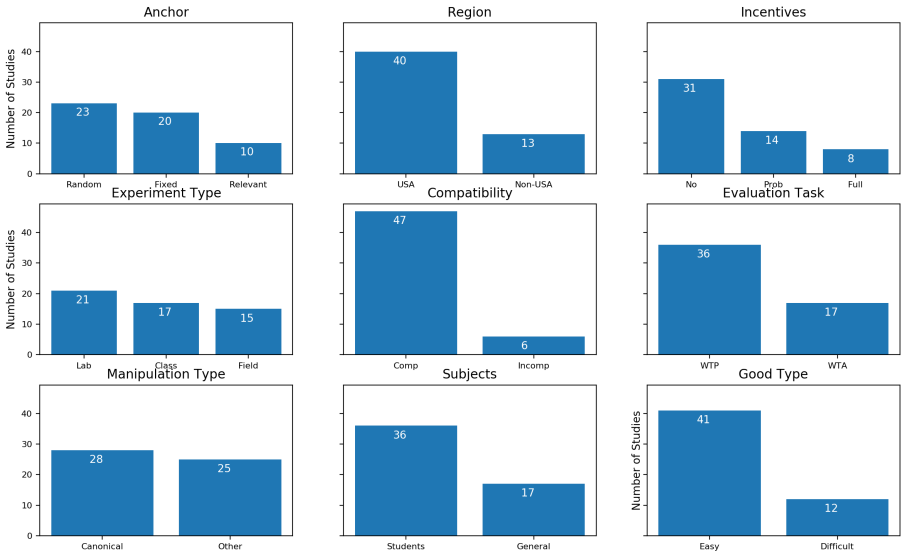
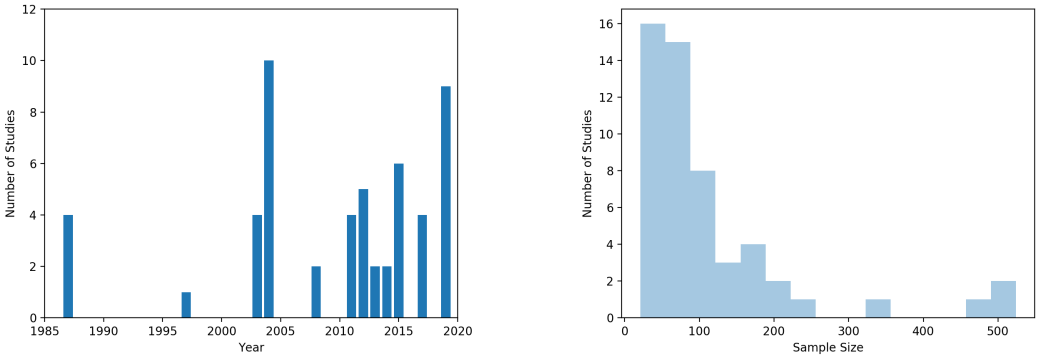


Figure 3.2: Number of studies against publication year and sample size, complete dataset (53 studies)



(a) (b)

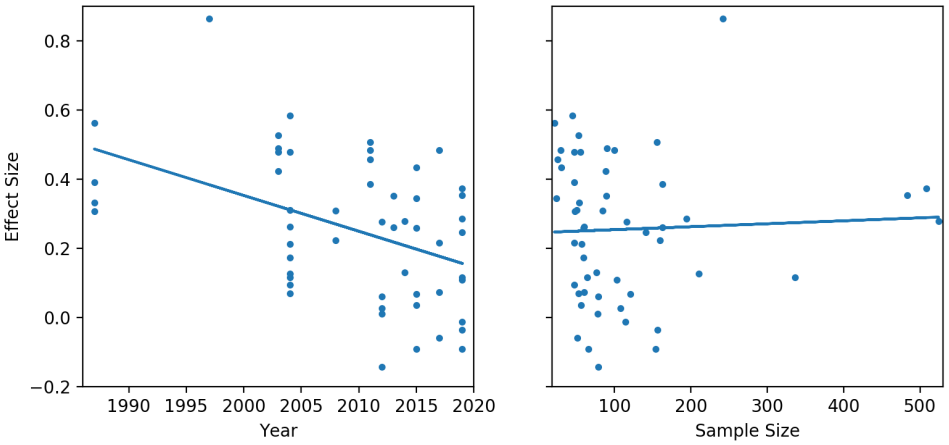
The reason is that the variance of r can be estimated using the following formula:

$$v_r = \frac{(1-r^2)^2}{n-1}, \tag{3.2}$$

where r is the sample correlation and n is the sample size. Since v_r highly depends on the correlation itself, this approximation is not recommended in meta-analysis. The z transformation avoids this problem, as the variance of z is:

$$v_z = \frac{1}{n-3}, \tag{3.3}$$

Figure 3.3: Effect size as a function of publication year and sample size, complete dataset (53 studies)



which is a simple and 'excellent approximation' (Cooper et al., 2009, p. 231). We followed the convention of first transforming the original r into z. Then, we perform the analysis using z. In the end, we convert the results expressed in terms of z back to r using the equation:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{3.4}$$

To combine estimates of effect size from different studies, we employ two standard sets of models. The fixed-effects model assumes that there is a single true (population) effect size for all studies, while the random-effects model assumes that there is variation in the true effect size between studies. The fixed-effect estimate of the overall correlation coefficient between anchor number and elicited valuation is 0.286, with 95% confidence interval [0.263, 0.309]. The results do not change much when we apply random effects analysis. The overall average effect size is 0.267, with 95% confidence interval [0.194, 0.338], and the estimate of between-study variance is 0.068. The results point to a moderate overall effect, smaller than the effects reported in early studies. The forest plots for both fixed and random meta-analysis are reported in Appendix B.1.

Importantly, we find substantial heterogeneity among studies. We performed a test of heterogeneity of the effect sizes and found a very large I-squared statistic equal to 88.2% (highly significant heterogeneity).[17] This indicates that differences across studies play a major role, and hence a deep examination of how these matter for the determination of the effect size is necessary.

Table 3.3: Sub-group random-effect estimates of the overall ES, complete dataset (53 studies)

Category	Overall ES	95% CI	# of Studies
Random	0.205	[0.136, 0.272]	23
Fixed	0.275	[0.101, 0.433]	20
Related	0.415	[0.328, 0.495]	10
WTP	0.273	[0.180, 0.360]	36
WTA	0.251	[0.144, 0.352]	17
Easy	0.270	[0.219, 0.318]	41
Difficult	0.256	[-0.018, 0.494]	12
Canonical	0.276	[0.167, 0.379]	28
Non-canonical	0.249	[0.166, 0.329]	25
Students	0.229	[0.162, 0.294]	36
General population	0.339	[0.185, 0.477]	17
Not incentivised	0.307	[0.200, 0.407]	31
Prob. incentivised	0.240	[0.136, 0.338]	14
Fully incentivised	0.163	[0.040, 0.281]	8
Lab	0.204	[0.127, 0.279]	21
Class	0.305	[0.183, 0.417]	17
Field	0.312	[0.144, 0.463]	15
Compatible	0.279	[0.201, 0.353]	47
Incompatible	0.141	[-0.011, 0.288]	6

3.3.3 Moderator Analyses

As we explained, we have coded a series of moderators that can be used as explanatory variables for the observed effect sizes. A sub-group analysis results based on the moderators is presented in Table 3.3. In addition, we calculate a meta-regression result based on those moderators. The model is as follows:

$$z_i = \alpha + X_i\beta + e_i + u_i, \tag{3.5}$$

where z_i is z transformation of r for study i,[18] X_i is a vector of coded moderators, $e_i \sim N(0, \sigma_i^2)$ captures the within-study variation, and $u_i \sim N(0, \tau^2)$ captures the between-study variation.[19]

The regression result is reported in Table 3.4. All variables are binary, and for all moderators we treat the most common category (see Figure 1) as the baseline, omitted variable. For instance, the baseline for the moderator 'type of anchor' is 'random anchor'

[17] I-squared measures the percentage of variation in effect sizes which is attributable to heterogeneity, rather than pure chance (Higgins & Thompson, 2002; Higgins, Thompson, Deeks & Altman, 2003).

[18] We have mentioned in Section 3.3.2 that r is not suitable for performing syntheses, thus here we use z in our regressions. However, for comparison purposes, and since the interpretation of r is more intuitive, we also report the results using r in Appendix B.2. These results are very similar in terms of the significance levels of the coefficients.

[19] We used the 'metareg' command in Stata. It is essentially variance-weighted least squares regression with a between-study variation u_i. Please note that the variance of e_i is known, since we have calculated it using equation 3.3 (and equation 3.2 if we regress r).

(category 1) and the baseline for the moderator 'type of experiment' is 'lab experiment' (category 1).

The estimated coefficients of our meta-regression in general have the expected sign. In particular, it seems that the presence of non-random anchors (either directly related to the goods or not) significantly increases the anchoring effect. This is consistent with our prior hypothesis about this, which has also been expressed in multiple studies. Moreover, selling tasks are weakly associated with a lower anchoring effect, consistent with the results by Simonson & Drolet (2004). Anchors that are scale/dimension-incompatible with the elicited valuation also seem to have a lower effect, as predicted by Strack & Mussweiler (1997). This is consistent with both the 'anchoring and adjustment' and the 'selective accessibility' theories of the anchoring effect. Compatible anchors are more likely to induce consumers to entertain the idea that the anchor may be a suitable answer to the elicitation question. This is a prerequisite for anchoring to work according to both theories. Experiments conducted outside the experimental laboratory generally yield stronger anchoring effects (but often insignificant), which could be attributed to the greater experimental control allowed by the lab. On the other hand, incentives do not seem to make much of a difference for the anchoring effect. This is consistent with prior findings. The lack of effect of incentives has been interpreted as providing evidence in favour of 'selective accessibility', rather than 'anchoring-and-adjustment'. The reason is that adjustment requires effort and should increase with incentives (resulting in lower anchoring).

Finally, the sign of the 'difficult' goods is surprisingly negative.[20] There are several plausible explanations for this. First, it could reflect the conceptual difficulty in defining what a 'difficult' good is and the practical limitations in successfully coding this construct. Secondly, it can be interpreted as an indirect rejection of the underlying hypothesis that consumers have 'an inventory of preferences' and report their WTP on the basis of this.[21] Finally, it could be driven by an alternative channel. For instance, if people know that the good is difficult to evaluate, they may become more cautious, thus less influenced by anchoring (Blankenship, Wegener, Petty, Detweiler-Bedell & Macy, 2008; Wegener, Petty, Detweiler-Bedell & Jarvis, 2001).

[20] The coefficient is insignificant when including Green et al. (1998), but it becomes significant when excluding Green et al. (1998) (see Appendix B.3).

[21] If this were the case, one would expect that 'easy' goods allow easy access to this 'inventory' and thus WTP is well-defined for such goods.

Table 3.4: Meta-regression on z, complete dataset (53 studies)

VARIABLES	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Fixed anchor	0.207**	0.199**	0.219**	0.220**	0.222**
	(0.0967)	(0.0930)	(0.0941)	(0.0938)	(0.0929)
Related anchor	0.357**	0.344***	0.357***	0.318***	0.332***
	(0.134)	(0.125)	(0.127)	(0.115)	(0.105)
Omitted var.: random anchor					
Prob. incentive	0.116	0.126	0.117	0.156	0.156
	(0.120)	(0.115)	(0.118)	(0.103)	(0.102)
Full incentive	-0.0529	-0.0615	-0.110	-0.116	-0.118
	(0.110)	(0.105)	(0.102)	(0.101)	(0.100)
Omitted var.: no incentive					
Class experiment	0.140	0.158	0.233**	0.224**	0.232**
	(0.125)	(0.108)	(0.0970)	(0.0959)	(0.0918)
Field experiment	0.275	0.280	0.316	0.249	0.322***
	(0.261)	(0.258)	(0.261)	(0.243)	(0.104)
Omitted var.: lab experiment					
WTA	-0.172*	-0.165*	-0.139*	-0.133	-0.134
	(0.0857)	(0.0820)	(0.0816)	(0.0810)	(0.0802)
Difficult	-0.0919	-0.0951	-0.108	-0.105	-0.109
	(0.0880)	(0.0864)	(0.0879)	(0.0875)	(0.0859)
Incompatible	-0.248*	-0.255*	-0.316**	-0.340**	-0.343**
	(0.144)	(0.140)	(0.136)	(0.132)	(0.130)
General Population	-0.0523	-0.0317	-0.0135	0.0716	
	(0.254)	(0.242)	(0.247)	(0.215)	
Non-canonical	-0.0851	-0.0751	-0.0707		
	(0.105)	(0.0983)	(0.100)		
2010 or later	-0.136	-0.129			
	(0.0945)	(0.0899)			
Non-USA	0.0350				
	(0.110)				
Constant	0.200	0.206	0.0874	0.0472	0.0452
	(0.143)	(0.140)	(0.115)	(0.0997)	(0.0988)
Observations	53	53	53	53	53

1. Standard errors in parentheses $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.
2. "Incompatible" represents scale or dimension incompatibility, and "Difficult" means that the value of the good is difficult to evaluate.

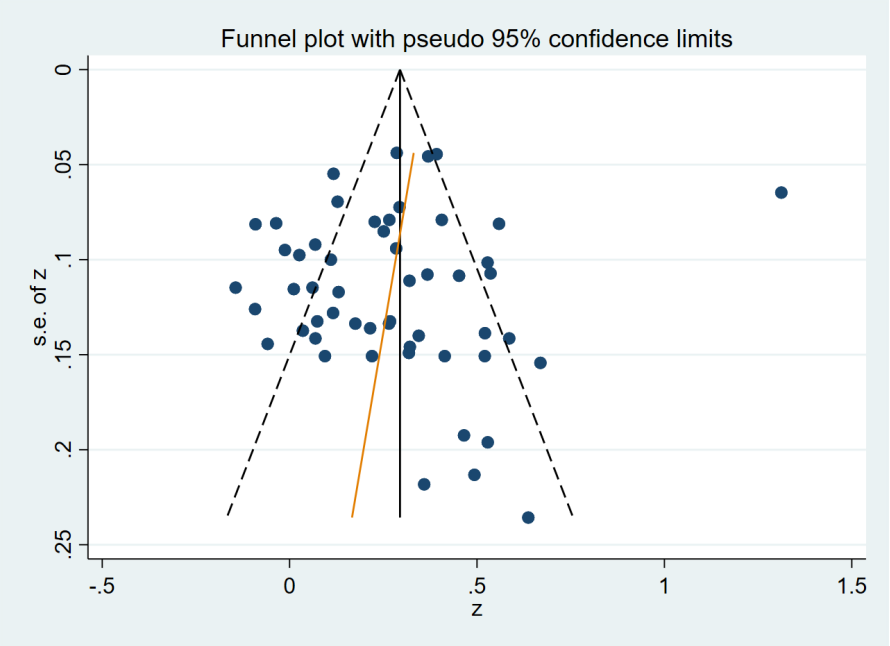
3.4 Robustness Checks

3.4.1 Publication Bias

It is well-known that the published literature is more likely to contain studies that report statistically significant effects, since researchers have incentives to place studies with insignificant results in the 'file drawer', and journal articles are more likely to publish studies with significant results. Also, it is possible that studies with interesting findings (significant effect sizes) are more likely to be accepted in more prestigious and

visible journals, which are easier for the meta-researcher to find. In general, the retrieved studies for a particular meta-analysis could only be a fraction of all relevant studies, and there is evidence suggesting that the hidden studies are systematically different from the retrieved ones (Song, Easterwood, Guilbody, Duley & Sutton, 2000; Dickersin, 2005). This is the publication bias problem, and it tends to distort the information available to the meta-researcher.[22]

Figure 3.4: Funnel plot with z, complete dataset (53 studies)



To assess this issue, we present the 'funnel plot' of our meta-analysis in Figure 3.4, noting that in the x-axis we have the z-transformation of r, and the y-axis presents the standard error of z (a measure of the precision of the effect size). The graph is asymmetric and reveals that small studies (studies with large s.e. of z) reporting small effect sizes are missing. This suggests the possibility of publication bias, but we need to be very careful when interpreting the funnel plot. Not only is the publication bias not the only possible cause of funnel plot asymmetry, but it is also the choice of the precision measurement that could significantly change the appearance of the plot. For instance, the asymmetry in the 'effect sizes against sample sizes' graph in Figure 3.3 is unnoticeable. Egger's linear regression test (Egger, Smith, Schneider & Minder, 1997) is thus used to statistically test the asymmetry. We regress the standardised effect size against a measure of result precision:

$$z_i/\sqrt{v_i} = \beta_0 + \beta_1(1/\sqrt{v_i}) + e_i, \tag{3.6}$$

where $e_i \sim N(0, s^2)$ and v_i is the sampling variance of study i. We are interested in the significance level of the intercept, since it is a measure of bias. The result shows that

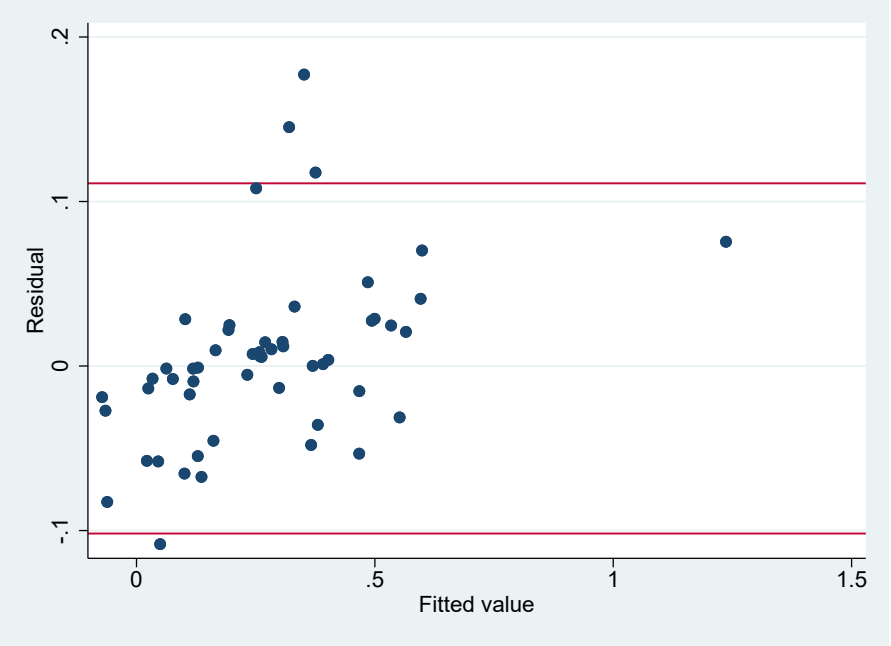
[22] Perhaps, 'dissemination bias' is a more accurate term to describe all the aforementioned types of bias. We follow the dominant convention and use 'publication bias' instead (Song et al., 2000).

the intercept is insignificant, with a p-value of 0.398, which makes us unable to conclude that there is significant asymmetry.

We carried out a rather comprehensive search of the literature. We retrieved some of the grey literature, such as working papers and unpublished manuscripts, which we believe may have helped in reducing the publication bias. Therefore, we tentatively conclude that publication bias is not a major problem in our analysis. However, it is possible that there are some small studies with small effect sizes left in the 'file drawer', in which case the overall effect size from the retrieved literature may be overestimated.

3.4.2 Other Potential Biases

Figure 3.5: Residual plot with z, complete dataset (53 studies)

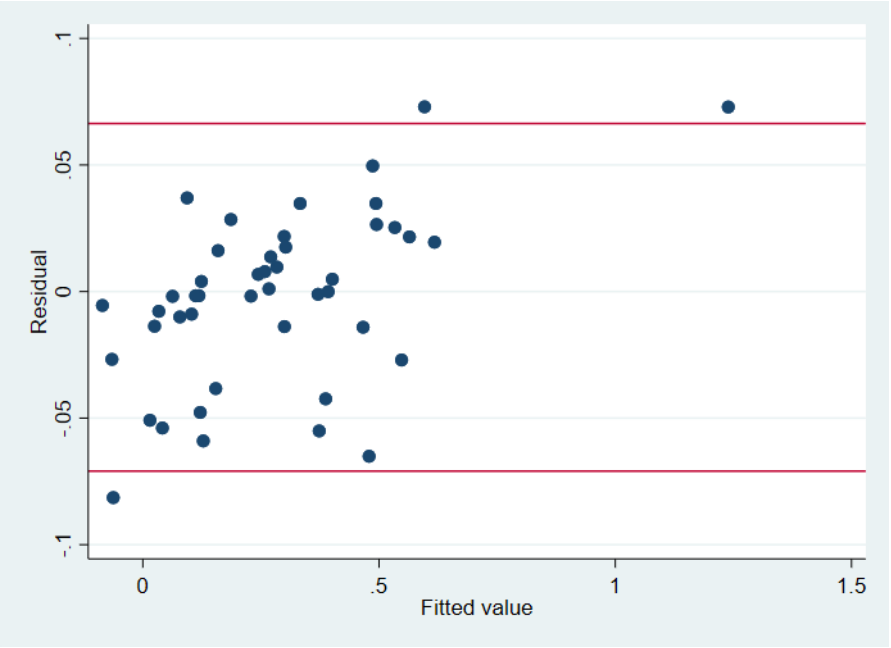


Several of our included articles contain multiple relevant studies. The analysis until this point has assumed that all studies are independent. In this section, we conduct some robustness analysis to examine the effect of possible 'research team-level bias'. Firstly, we follow the suggestion in the Cochrane Handbook (Higgins, Thomas, Chandler, Cumpston, Li, Page & Welch, 2019) to address this possible concern. For all articles with multiple studies, we select one study from each article at random and exclude others. This leaves us with a reduced dataset which contains 24 studies.[23] The fixed-effect estimate of the overall effect size is 0.311 with 95% confidence interval [0.281, 0.342], and random-effect estimate of the overall effect size is 0.300 with 95% confidence interval [0.173, 0.416]. The combined estimate of the effect size are slightly larger than of the complete dataset. In terms of the meta-regression, now only the coefficients of

[23]The selection is random, and Appendix B.4 provides a general description of the selected studies. Figure B.5 shows that in the reduced dataset a level of heterogeneity remains. Figure B.7 plots effect sizes against time and sample size and the tendency is similar to the complete dataset. See Appendix B.5 for detailed information on the selected studies.

'Full incentive' and 'Incompatibility' are significant as we have few observations, while the signs of the coefficients are consistent with the main analysis (see Appendix B.4 for details).

Figure 3.6: Residual plot with z, reduced dataset (45 studies)



Also, we checked whether our main regression results are affected by the 'article-level' effect by looking at the residual-fitted value plot in Figure 3.5. Note that the fitted value is $\alpha + \widehat{X\beta} + e$ (in model 3.5, the variance of e_i is known) and the red lines indicate ± 2 standard deviations from the mean. The residuals between the two red lines show no clear pattern (probably a weak positive correlation).[24] Among the five outliners, the one below -0.1 belongs to Brzozowicz et al. (2017), and it is worth noticing that the other four between (about) 0.1 and 0.2 belongs to only two articles: Adaval & Wyer (2011) and Dogerlioglu-Demir & Koçaş (2015). This might indicate that some 'article-level' effect driven by some unobservable characteristics of those studies are not coded into our moderators. To examine whether this drives our results, we conduct robustness checks where we leave out all studies contained in those three articles. We perform the regression with the remaining 45 studies, and the result is presented in Table 3.5. It shows similar results to Table 3.4, and the new residual-fitted value plot shows less tendency of heteroscedasticity (see Figure 3.6). It appears that the 'article level' effect does not drive the results of our main regression analysis.

[24] The graph suggests heteroscedasticity, which points to the possibility of omitted variables. However, we cannot add new variables into our regression as we have low number of obeservations. Also, we need to avoid the 'data dredging' issue mentioned in Section 3.2.

Table 3.5: Meta-regression on z, reduced dataset (45 studies)

VARIABLES	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Fixed anchor	0.129	0.145	0.195*	0.215**	0.217**
	(0.104)	(0.101)	(0.103)	(0.101)	(0.1000)
Related anchor	0.354**	0.386***	0.391***	0.326***	0.344***
	(0.137)	(0.128)	(0.135)	(0.118)	(0.108)
Omitted var.: random anchor					
Prob. incentive	0.154	0.128	0.125	0.180	0.180
	(0.123)	(0.116)	(0.122)	(0.109)	(0.108)
Full incentive	-0.0661	-0.0506	-0.122	-0.127	-0.129
	(0.115)	(0.111)	(0.111)	(0.111)	(0.110)
Omitted var.: no incentive					
Class experiment	0.191	0.144	0.246**	0.220**	0.231**
	(0.130)	(0.112)	(0.105)	(0.102)	(0.0971)
Field experiment	0.383	0.369	0.397	0.279	0.369***
	(0.269)	(0.265)	(0.278)	(0.251)	(0.113)
Omitted var.: lab experiment					
WTA	-0.164*	-0.180**	-0.141*	-0.133	-0.135
	(0.0845)	(0.0807)	(0.0825)	(0.0823)	(0.0815)
Difficult	-0.0642	-0.0517	-0.103	-0.116	-0.122
	(0.105)	(0.103)	(0.105)	(0.105)	(0.102)
Incompatible	-0.389**	-0.371**	-0.391**	-0.400**	-0.400**
	(0.181)	(0.177)	(0.187)	(0.188)	(0.186)
General Population	-0.0649	-0.114	-0.0532	0.0880	
	(0.263)	(0.251)	(0.261)	(0.218)	
Non-canonical	-0.133	-0.156	-0.115		
	(0.119)	(0.113)	(0.118)		
2010 or later	-0.173*	-0.191*			
	(0.0985)	(0.0944)			
Non-USA	-0.0851				
	(0.119)				
Constant	0.281*	0.268*	0.0806	0.0265	0.0237
	(0.149)	(0.146)	(0.120)	(0.106)	(0.105)
Observations	45	45	45	45	45

1. Standard errors in parentheses $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.
2. "Incompatible" represents scale or dimension incompatibility, and "Difficult" means that the value of the good is difficult to evaluate.

3.5 Conclusion

The anchoring bias in consumers' assessments of Willingness-to-Pay is generally considered to induce a very robust and large effect. We have conducted a research synthesis on the importance and determinants of anchoring effects on statements of economic valuation. We retrieved 53 studies from 24 articles using an exhausting search of the literature. We obtained an effect (correlation coefficient between anchor and target item) of moderate size. This is generally smaller than the effects revealed in early studies

of the phenomenon. Our analysis also uncovers substantial heterogeneity of the effect size. To address this, we performed a meta-regression after coding for moderators that could drive the effect. We find that experiments conducted in the classroom are likely to find larger anchoring effects, as do studies where the anchor is possibly informative about the true quality of the evaluated item.

Our findings have some important implications, because they show that anchoring might not be as strong and robust as considered so far. The lack of ubiquitous strong anchoring effects (especially when anchors are clearly random) imply that well-defined and constant preferences may still be a good approximation. In terms of theory, our results seem to provide support for the 'selective accessibility' model of anchoring.

Appendix B

B.1 Forest plots with z, complete dataset

Figure B.1: Fixed effect model, complete dataset (53 studies)

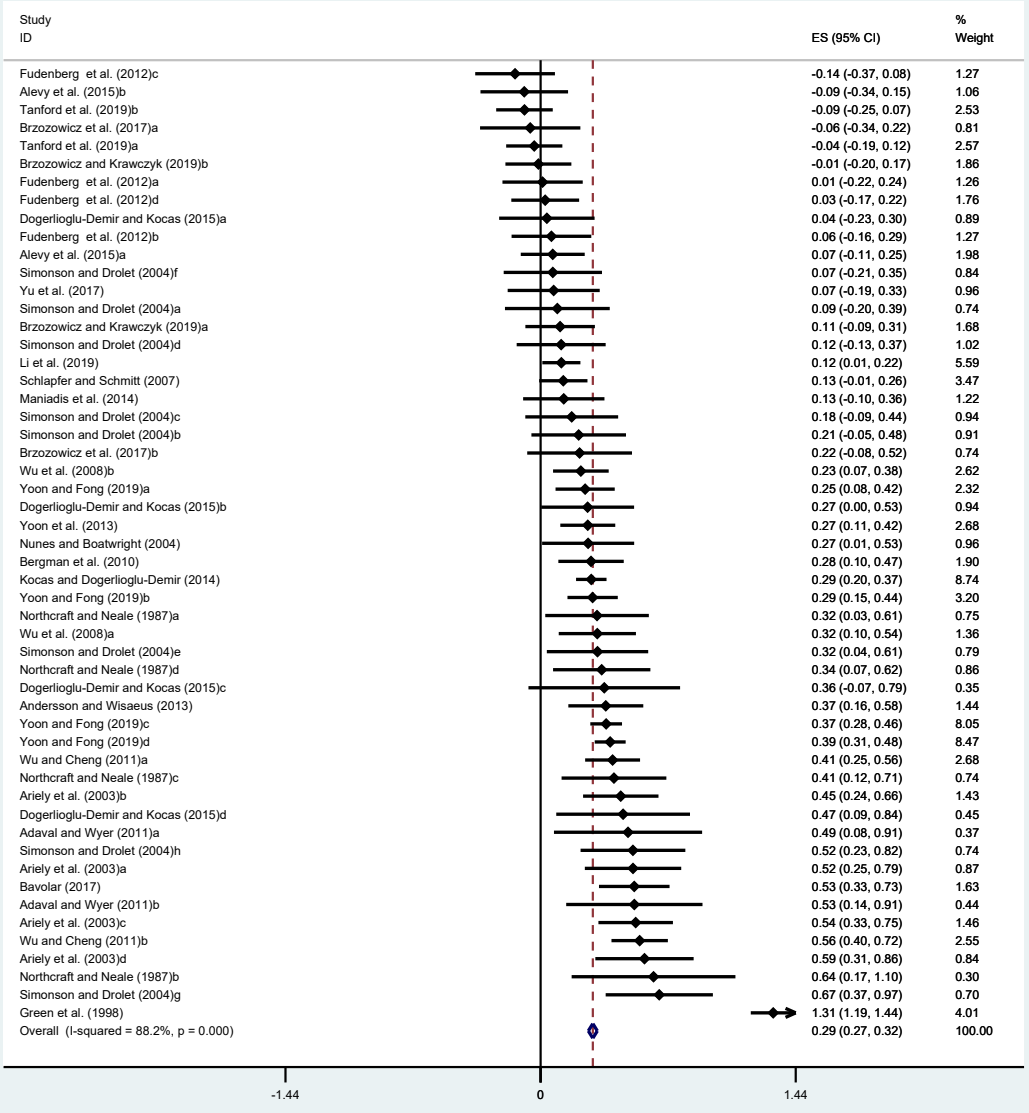
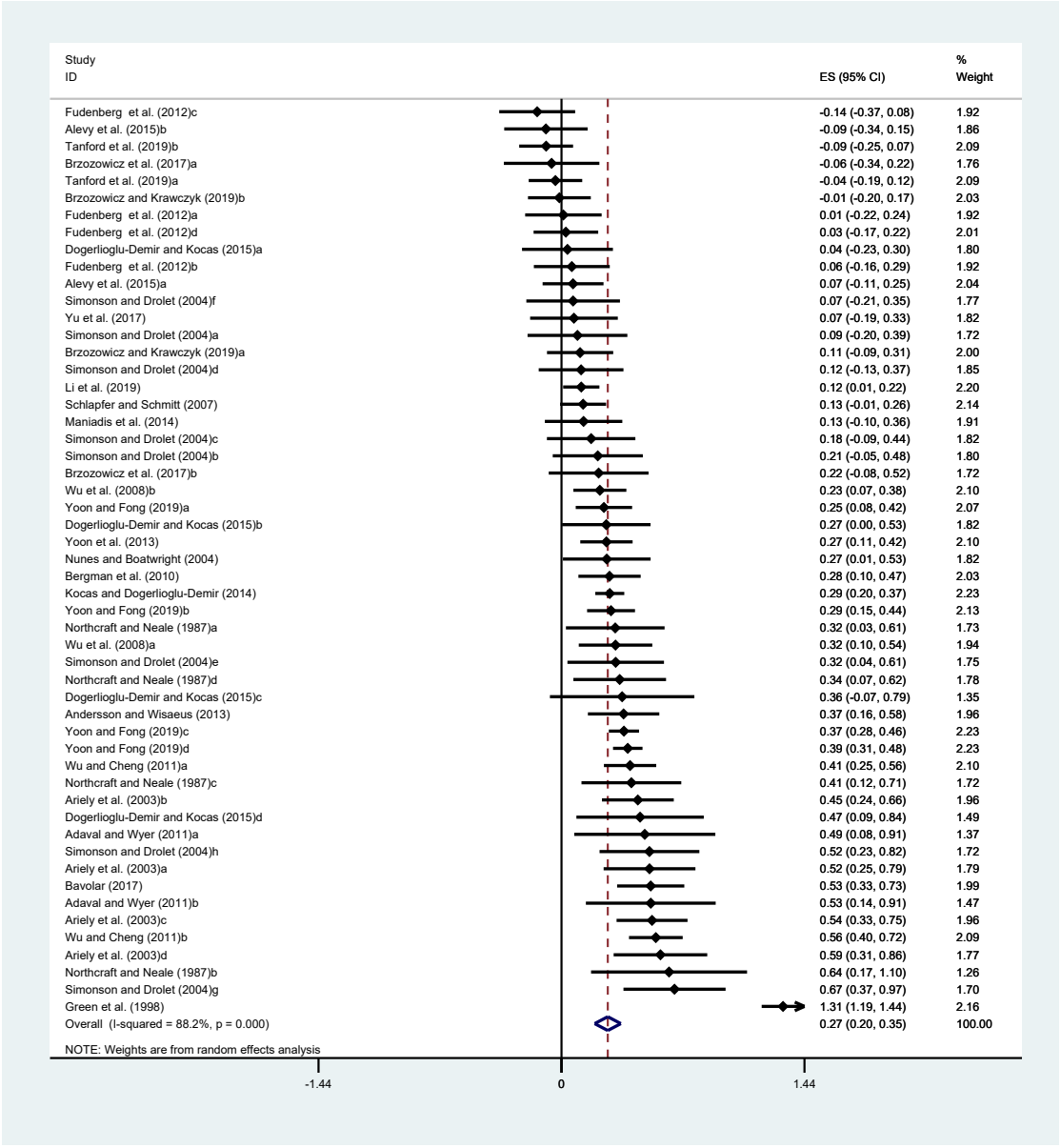


Figure B.2: Random effect model, complete dataset (53 studies)



B.2 Meta-regression on r, complete dataset

Table B.1: Meta-regression on r, complete dataset (53 studies)

VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed anchor	0.155**	0.154**	0.172**	0.173**	0.174**
	(0.0761)	(0.0734)	(0.0746)	(0.0744)	(0.0737)
Related anchor	0.312***	0.313***	0.322***	0.294***	0.301***
	(0.107)	(0.0989)	(0.101)	(0.0913)	(0.0841)
Omitted var.: random anchor					
Prob. incentive	0.112	0.112	0.105	0.134	0.134
	(0.0960)	(0.0917)	(0.0941)	(0.0825)	(0.0818)
Full incentive	-0.0491	-0.0489	-0.0884	-0.0937	-0.0948
	(0.0882)	(0.0842)	(0.0821)	(0.0815)	(0.0807)
Omitted var.: no incentive					
Class experiment	0.141	0.141	0.200**	0.193**	0.197***
	(0.0980)	(0.0848)	(0.0771)	(0.0763)	(0.0729)
Field experiment	0.251	0.252	0.276	0.228	0.264***
	(0.203)	(0.201)	(0.205)	(0.191)	(0.0819)
Omitted var.: lab experiment					
WTA	-0.144**	-0.144**	-0.121*	-0.117*	-0.117*
	(0.0686)	(0.0655)	(0.0655)	(0.0650)	(0.0644)
Difficult	-0.111	-0.111	-0.120*	-0.118*	-0.120*
	(0.0695)	(0.0682)	(0.0700)	(0.0698)	(0.0685)
Incompatible	-0.208*	-0.208*	-0.255**	-0.273**	-0.275**
	(0.113)	(0.110)	(0.109)	(0.105)	(0.104)
General Population	-0.0441	-0.0451	-0.0270	0.0354	
	(0.200)	(0.190)	(0.195)	(0.169)	
Non-canonical	-0.0560	-0.0566	-0.0522		
	(0.0838)	(0.0784)	(0.0806)		
2010 or later	-0.104	-0.105			
	(0.0740)	(0.0704)			
Non-USA	-0.000180				
	(0.0877)				
Constant	0.197*	0.198*	0.0992	0.0695	0.0686
	(0.113)	(0.111)	(0.0919)	(0.0795)	(0.0787)
Observations	53	53	53	53	53

1. Standard errors in parentheses $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.
2. "Incompatible" represents scale or dimension incompatibility, and "Difficult" means that the value of the good is difficult to evaluate.

B.3 Meta-analytic results, reduced dataset with (Green et al., 1998) being excluded (52 studies)

- Fixed effect model: overall effect size is 0.247 with 95% confidence interval [0.222, 0.271]
- Random effect model: overall effect size is 0.241 with 95% confidence interval [0.190, 0.290]

- Estimate of between-study variance Tau-squared = 0.024
- I-squared (variation in effect size attributable to heterogeneity) = 72.3%

Table B.2: Sub-group random-effect estimates of the overall ES, reduced dataset (52 studies)

Category	Overall ES	95% CI	# of Studies
Random	0.205	[0.136, 0.272]	23
Fixed	0.199	[0.114, 0.281]	19
Related	0.415	[0.328, 0.495]	10
WTP	0.238	[0.180, 0.295]	35
WTA	0.251	[0.144, 0.352]	17
Easy	0.270	[0.219, 0.318]	41
Difficult	0.141	[0.028, 0.253]	11
Canonical	0.236	[0.172, 0.299]	27
Non-canonical	0.249	[0.166, 0.329]	25
Students	0.229	[0.162, 0.294]	36
General population	0.265	[0.186, 0.340]	16
Not incentivised	0.263	[0.198, 0.326]	30
Prob. incentivised	0.240	[0.136, 0.338]	14
Fully incentivised	0.163	[0.040, 0.281]	8
Lab	0.204	[0.127, 0.279]	21
Class	0.305	[0.183, 0.417]	17
Field	0.235	[0.156, 0.311]	14
Compatible	0.252	[0.200, 0.302]	46
Incompatible	0.141	[-0.011, 0.288]	6

Table B.3: Meta-regression on z, reduced dataset (52 studies)

VARIABLES	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Fixed anchor	0.0665	0.100	0.107	0.107*	0.107*
	(0.0646)	(0.0640)	(0.0638)	(0.0631)	(0.0621)
Related anchor	0.294***	0.347***	0.351***	0.334***	0.339***
	(0.0907)	(0.0863)	(0.0862)	(0.0777)	(0.0720)
Omitted var.: random anchor					
Prob. incentive	0.157*	0.122	0.119	0.135*	0.135*
	(0.0811)	(0.0796)	(0.0795)	(0.0702)	(0.0693)
Full incentive	-0.0513	-0.0247	-0.0434	-0.0473	-0.0477
	(0.0734)	(0.0730)	(0.0702)	(0.0690)	(0.0681)
Omitted var.: no incentive					
Class experiment	0.233***	0.165**	0.194***	0.189***	0.191***
	(0.0835)	(0.0728)	(0.0664)	(0.0649)	(0.0620)
Field experiment	0.211	0.202	0.204	0.175	0.203***
	(0.183)	(0.186)	(0.186)	(0.174)	(0.0695)
Omitted var.: lab experiment					
WTA	-0.0917	-0.118**	-0.102*	-0.0994*	-0.0998*
	(0.0600)	(0.0583)	(0.0559)	(0.0550)	(0.0543)
Difficult	-0.200***	-0.180***	-0.187***	-0.187***	-0.189***
	(0.0600)	(0.0604)	(0.0600)	(0.0594)	(0.0581)
Incompatible	-0.160*	-0.144	-0.159*	-0.168*	-0.168*
	(0.0919)	(0.0944)	(0.0932)	(0.0902)	(0.0889)
General Population	0.0524	-0.0240	-0.00971	0.0271	
	(0.181)	(0.176)	(0.175)	(0.155)	
Non-canonical	0.00132	-0.0363	-0.0316		
	(0.0709)	(0.0679)	(0.0678)		
2010 or later	-0.0243	-0.0579			
	(0.0634)	(0.0610)			
Non-USA	-0.120				
	(0.0766)				
Constant	0.178*	0.164*	0.109	0.0927	0.0925
	(0.0932)	(0.0957)	(0.0772)	(0.0674)	(0.0664)
Observations	52	52	52	52	52

1. Standard errors in parentheses $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

2. "Incompatible" represents scale or dimension incompatibility, and "Difficult" means that the value of the good is difficult to evaluate.

Figure B.3: Funnel plot with z, reduced dataset (52 studies)

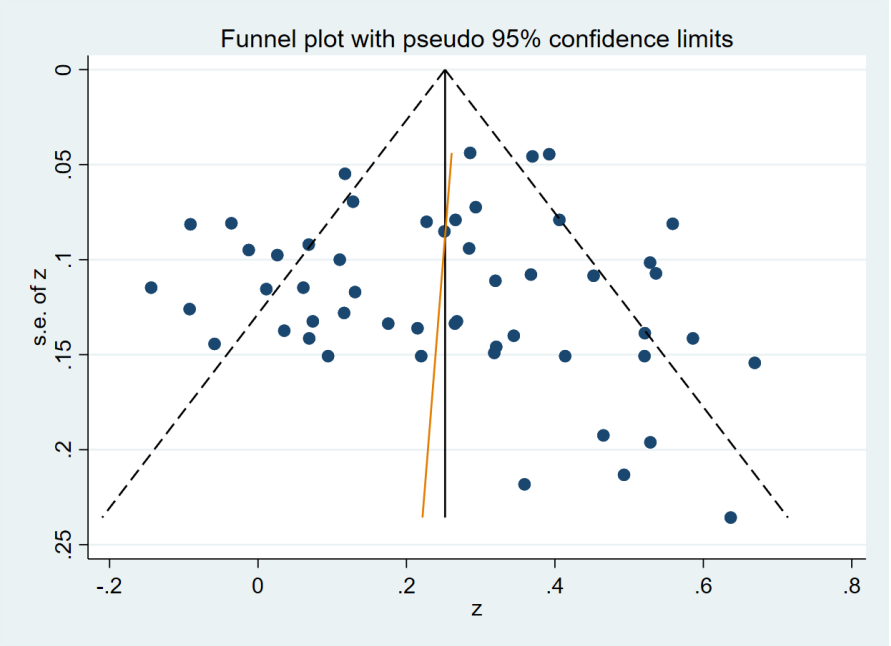
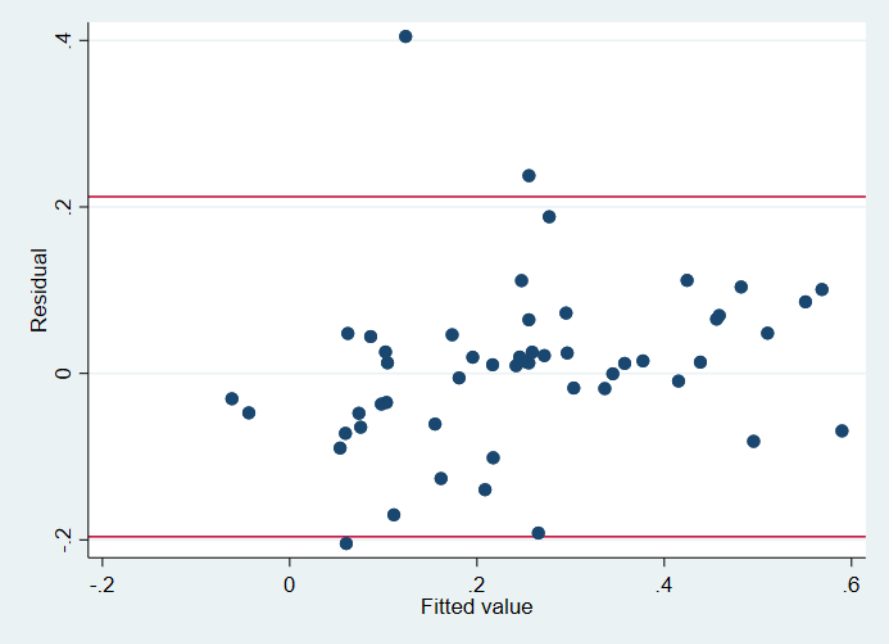


Figure B.4: Residual plot with z, reduced dataset (52 studies)



B.4 Descriptions and Meta-analytic results, reduced dataset (24 studies)

Figure B.5: Summary of number of studies in each category, reduced dataset (24 studies)

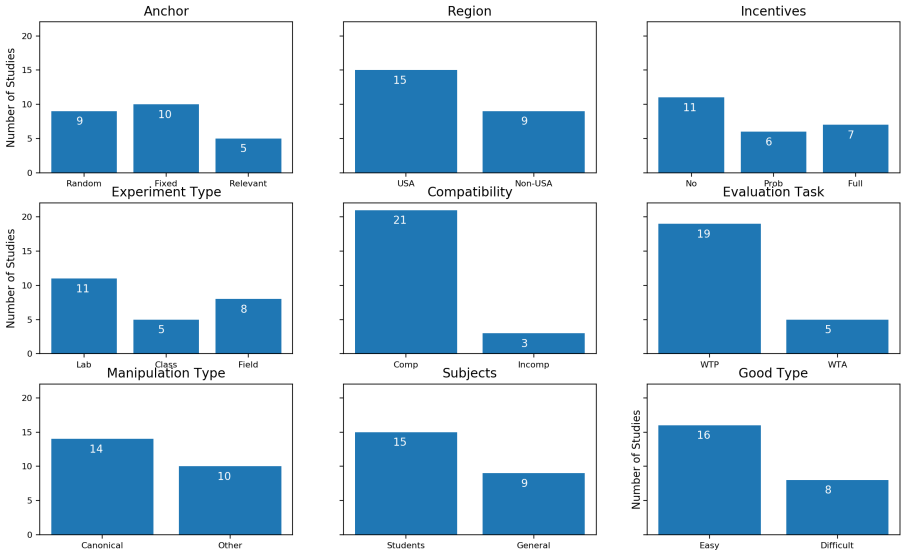
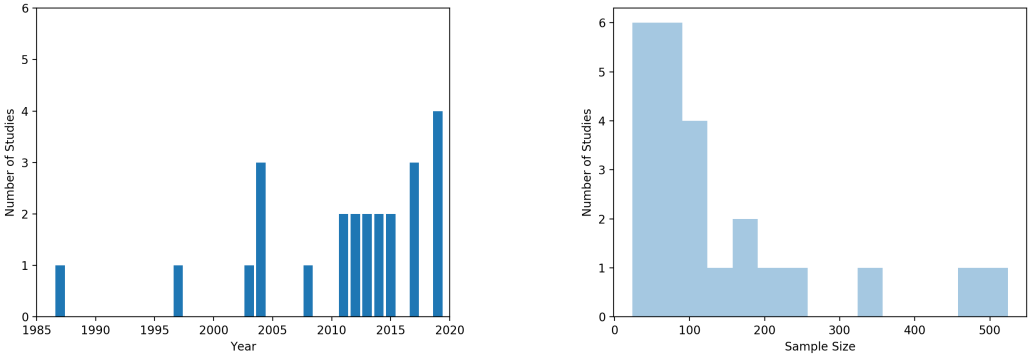


Figure B.6: Number of studies against publication year and sample size, reduced dataset (24 studies)



(a) (b)

Figure B.7: Effect size as a function of publication year and sample size, reduced dataset (24 studies)

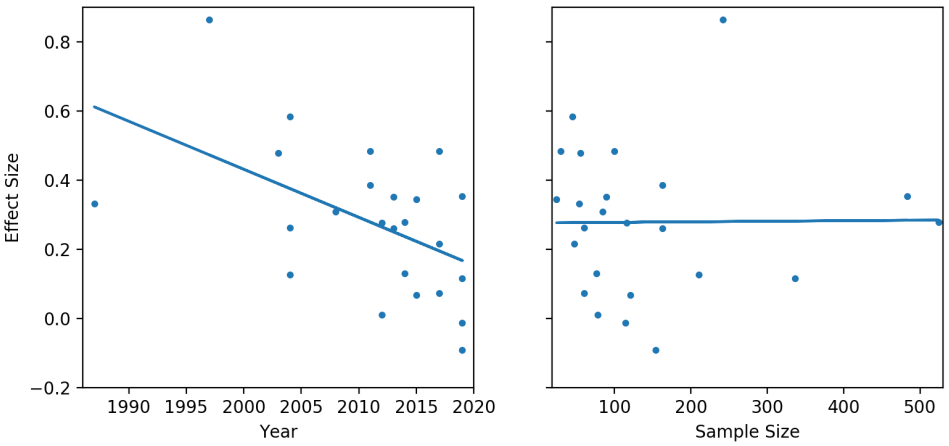


Table B.4: Meta-regression on z, reduced dataset (24 studies)

VARIABLES	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Fixed anchor	-0.00393	-0.00938	0.0400	0.0978	0.0992
	(0.251)	(0.240)	(0.246)	(0.222)	(0.217)
Related anchor	0.00152	-0.0163	0.0372	0.0223	0.0342
	(0.292)	(0.265)	(0.271)	(0.266)	(0.258)
Omitted var.: random anchor					
Prob. incentive	-0.294	-0.293	-0.262	-0.165	-0.178
	(0.331)	(0.317)	(0.328)	(0.281)	(0.272)
Full incentive	-0.209	-0.223	-0.251	-0.274	-0.293*
	(0.200)	(0.177)	(0.183)	(0.176)	(0.164)
Omitted var.: no incentive					
Class experiment	0.233	0.251	0.295	0.298	0.330
	(0.245)	(0.215)	(0.220)	(0.217)	(0.192)
Field experiment	0.0371	0.0471	0.0320	0.00860	0.164
	(0.493)	(0.471)	(0.488)	(0.479)	(0.211)
Omitted var.: lab experiment					
WTA	-0.197	-0.186	-0.191	-0.162	-0.138
	(0.189)	(0.172)	(0.178)	(0.169)	(0.152)
Difficult	0.0987	0.0946	0.0909	0.0491	0.0415
	(0.184)	(0.173)	(0.180)	(0.162)	(0.157)
Incompatible	-0.577*	-0.593**	-0.607**	-0.636**	-0.658**
	(0.274)	(0.253)	(0.263)	(0.254)	(0.241)
General Population	-0.0629	-0.0508	0.0666	0.154	
	(0.473)	(0.449)	(0.455)	(0.424)	
Non-canonical	-0.135	-0.117	-0.129		
	(0.244)	(0.210)	(0.218)		
2010 or later	-0.197	-0.195			
	(0.162)	(0.154)			
Non-USA	0.0342				
	(0.179)				
Constant	0.661	0.672*	0.461	0.357	0.362
	(0.380)	(0.362)	(0.335)	(0.281)	(0.273)
Observations	24	24	24	24	24

1. Standard errors in parentheses $***p < 0.01$, $**p < 0.05$, $*p < 0.1$.

2. "Incompatible" represents scale or dimension incompatibility, and "Difficult" means that the value of the good is difficult to evaluate.

B.5 Coding

Table B.5: Coding

Study	AT	TT	GT	MT	SP	IT	ET	CP	ES
Adaval and Wyer (2011)a	1	0	0	1	0	0	0	0	0.457
Adaval and Wyer (2011)b ∗	1	0	1	1	0	0	0	0	0.484
Alevy et al. (2015)a ∗	0	1	0	0	1	2	2	0	0.068
Alevy et al. (2015)b	0	1	1	0	1	2	2	0	-0.092
Andersson and Wisaeus (2013)	0	0	0	0	0	1	0	0	0.352
Ariely et al. (2003)a ∗	0	0	0	0	0	1	1	0	0.479
Ariely et al. (2003)b	1	1	0	0	0	1	1	0	0.424
Ariely et al. (2003)c	0	1	0	0	0	1	1	0	0.490
Ariely et al. (2003)d	1	1	0	0	0	1	1	0	0.527
Bavolar (2017)	2	0	1	1	0	2	1	0	0.484
Bergman et al. (2010)	0	0	0	0	0	1	0	0	0.277
Brzozowicz and Krawczyk (2019)a	1	0	1	0	0	0	0	0	0.110
Brzozowicz and Krawczyk (2019)b ∗	1	0	1	0	0	1	0	0	-0.012
Brzozowicz et al. (2017)a	1	0	0	1	0	0	0	0	-0.059
Brzozowicz et al. (2017)b ∗	1	0	0	1	0	2	0	0	0.216
Dogerlioglu-Demir and Kocas (2015)a	1	0	0	1	1	0	2	1	0.035
Dogerlioglu-Demir and Kocas (2015)b	1	0	0	1	1	0	2	1	0.259
Dogerlioglu-Demir and Kocas (2015)c ∗	1	0	0	1	0	0	1	1	0.344
Dogerlioglu-Demir and Kocas (2015)d	1	0	0	1	0	0	1	1	0.434
Fudenberg et al. (2012)a ∗	0	1	0	0	0	1	0	0	0.011
Fudenberg et al. (2012)b	0	1	0	0	0	1	0	0	0.061
Fudenberg et al. (2012)c	0	0	0	0	0	1	0	0	-0.143
Fudenberg et al. (2012)d	0	1	0	0	0	1	0	0	0.026
Green et al. (1998)	1	0	1	0	1	0	2	0	0.865
Kocas and Dogerlioglu-Demir (2014)	1	0	0	0	1	0	2	0	0.278
Li et al. (2019)	0	0	1	1	1	2	2	0	0.117
Maniadis et al. (2014)	1	1	0	0	0	2	0	0	0.130
Northcraft and Neale (1987)a	2	1	1	1	1	0	2	0	0.308
Northcraft and Neale (1987)b	2	1	0	1	1	0	2	0	0.563
Northcraft and Neale (1987)c	2	1	0	1	1	0	2	0	0.392
Northcraft and Neale (1987)d ∗	2	1	1	1	1	0	2	0	0.332
Nunes and Boatwright (2004)	0	0	0	1	1	2	2	0	0.262
Schlapfer and Schmitt (2007)	1	0	1	0	1	0	2	1	0.127
Simonson and Drolet (2004)a	0	1	0	0	0	0	1	0	0.094
Simonson and Drolet (2004)b	0	1	0	1	0	0	1	0	0.212
Simonson and Drolet (2004)c	0	1	0	1	0	0	1	0	0.174
Simonson and Drolet (2004)d	0	0	0	0	0	0	1	0	0.115
Simonson and Drolet (2004)e	0	0	0	1	0	0	1	0	0.310
Simonson and Drolet (2004)f	0	0	0	1	0	0	1	0	0.069
Simonson and Drolet (2004)g ∗	2	1	0	0	1	0	1	0	0.584
Simonson and Drolet (2004)h	2	0	0	0	1	0	1	0	0.478
Tanford et al. (2019)a	1	0	1	1	0	0	1	0	-0.036
Tanford et al. (2019)b ∗	1	0	1	1	0	0	1	1	-0.091
Wu and Cheng (2011)a ∗	2	0	0	1	0	0	0	0	0.385
Wu and Cheng (2011)b	2	0	0	1	0	0	0	0	0.507
Wu et al. (2008)a ∗	1	0	0	0	0	0	0	0	0.309

Table B.5 continued from previous page

Wu et al. (2008)b	1	0	0	1	0	0	0	0	0.223
Yoon and Fong (2019)a	0	0	0	0	0	1	0	0	0.246
Yoon and Fong (2019)b	0	0	0	0	0	1	0	0	0.285
Yoon and Fong (2019)c $*$	0	0	0	0	1	0	2	0	0.354
Yoon and Fong (2019)d	0	0	0	0	1	0	2	0	0.373
Yoon et al. (2013)	0	0	0	0	0	1	0	0	0.260
Yu et al. (2017)	2	0	0	1	0	2	0	0	0.074

Note: 1. see Table 3.1 for abbreviations and variable definitions; 2. $*$ indicates the study is randomly selected for our analysis in Section 3.4.2.

Chapter 4

Can Biased Polls Distort Electoral Results? Evidence from the Lab and Field

4.1 Introduction

The rise of populism in western democracies over the last few years has changed the political landscape, upsetting political balances that survived for decades and bringing new forces into the forefront. Some academics attribute the phenomenon to economic factors (Funke, Schularick & Trebesch, 2016; Guriev, 2018), while others to recent developments in traditional and social media (DellaVigna & Kaplan, 2007; Petrova, 2008; Boxell, Gentzkow & Shapiro, 2017), or to cultural factors (Oesch, 2008; Van Hauwaert & Van Kessel, 2018).[1] Regardless of the origins of its resurgence, populism has ramifications for both economic policy (Kaltwasser, 2018) and political stability. A key aspect of modern populism is distrust in democratic institutions, of which the most fundamental is elections under free media.

The role of voting-intention polls, in particular, has been under heavy criticism. Questions have been raised about the reliability of polls (Shirani-Mehr, Rothschild, Goel & Gelman, 2018) and their effects on democratic elections. A key reason for such scepticism is widespread perception of poor predictive performance of polls in recent high-profile elections, most notably the 2016 US presidential election and the UK general elections of 2015 and 2017.[2] Some prominent politicians, such as Lord Foulkes in the UK and Ron Paul in the US, have even claimed that polls may be manipulated by status-quo groups in an attempt to cling to power.[3] If people perceive polls to be biased in favour

[1] For broader treatments of populism, see also the papers by Boeri, Mishra, Papageorgiou & Spilimbergo (2018), Inglehart & Norris (2017) and Rodrik (2018). For the role of social media in the proliferation of fake news see Allcott & Gentzkow (2017).

[2] See also Whiteley (2016) on the reasons behind the failure of polls to predict the 2015 general election in the UK.

[3] For general economic models of such manipulation see Maniadis (2014) and Cipullo & Reslow (2019).

of a candidate or a party, this perception may erode trust in democratic institutions and reinvigorate populist agendas.

Thus, it is crucial to examine whether criticisms of polls, such as the ones presented above, are unfounded, or whether polls have the potential to skew election results in the current political environment. If the feedback that the public receives from opinion polls is not representative of the true preferences of the electorate, one may worry that this could distort the democratic process. For instance, in a two-party election race, imagine that poll results showing the left party ahead are more likely to be revealed to the public than poll results showing this party trailing. Then, a critical question arises: would this systematic bias in exposure to poll results affect the electoral race? How significant is such an effect and how does it depend on what voters know about the bias? In this paper, we first show that this type of biased exposure to the results of polls may arise naturally from the dynamics of modern communication (social media). We then examine the causal effect of such biased exposure on elections. We use the experimental approach and a series of robustness checks and find that biased exposure consistently leads to meaningful and systematic changes in election outcomes.

Even leaving aside the possibility of conscious manipulation, several plausible mechanisms could generate systematic bias in the feedback that citizens receive about the results of voting-intention polls (Sturgis, Baker, Callegaro, Fisher, Green, Jennings, Kuha, Lauderdale & Smith, 2016). First of all, pollsters have methodological flexibility similar to other empirical scientists (Ioannidis, 2005) and if they have strong priors about who is leading, they may choose methods that verify these priors (for example, turnout adjustments). Moreover, the traditional media reveal poll results selectively, either to pander to the expectations of their audience (Gentzkow & Shapiro, 2010) or to simply make interesting news (Larsen & Fazekas, 2019). Finally, the voters themselves may propagate results in a biased manner, especially via social media. Our objective in this paper is to first empirically substantiate such biased feedback and then examine its causal effects on election outcomes.

It is difficult to find appropriate data to substantiate most of the mechanisms described above. For this reason, in Section 4.3 we shall provide empirical evidence regarding only the last channel, which mediates communication and may lead to biased propagation of poll results: online publics. We use Twitter data from the US and the UK and examine econometrically how the patterns of retweeting of news about poll results are affected by the results themselves. We find systematic biases in the manner in which poll results are propagated via social media. In particular, there is a fundamental asymmetry between parties in the pattern of propagation. In our data, 'good news' about the popularity of conservative parties seem to receive less propagation than 'bad news', whereas the opposite is true about liberal parties.

After having empirically established the existence of biased exposure of the public to poll results, we turn to an examination of its consequences. Given the great difficulty of using observational data not only to measure the degree of bias in polls but also to examine the electoral consequences of this, we take an experimental approach

to address the question. This allows us to control voters' information both at the poll and at the election stage. We focus on an environment of two-party elections and we postulate a straightforward 'biased rule' according to which the revelation of poll results takes place. In particular, only the most 'favourable' results for a particular party (the 'favoured party') are revealed. This censored rule captures the essence of the idea that propagation of poll results to the public depends on the results themselves. This information pattern could ensue from any of the reasons discussed earlier: conscious manipulation, pollsters' priors that the favoured party is ahead (coupled with methodological flexibility), traditional media wishing to match the public's expectations, the biased way social media propagate poll outcomes, etc.

In our experiments, we observe the outcome of fifteen electoral races between the same two parties (we call them parties K and J) who field different candidates every time. The two candidates differ in their 'valence', and the exact valences are known to only some participants (the 'informed voters'). 'Uninformed voters' are only told the statistical distribution out of which the valences were drawn. Before each election, five voting-intention polls are generated by randomly sampling participants. In this manner, polls allow informed voters to provide a noisy signal regarding the valence of the two candidates.

In Experiment 1 (E1), we start by comparing a biased regime – where the results of only the *two polls most favourable for one candidate* (the candidate of party K, or simply *candidate K*) are revealed – to a natural control setting, where *all five polls* are revealed. In addition, we conduct two robustness checks. In Experiment 2 (E2), the control setting entails revealing the results of *two randomly selected polls*, rather than all five polls. Finally, in Experiment 3 (E3), we keep the same control condition as E1, but in the treatment condition participants are informed beforehand about the (non-random) rule for selecting the two polls.

If a party's popularity is systematically 'inflated' in the polls, does this result in an electoral advantage for that party? Our results suggest that this is indeed the case. Both in terms of the number of rounds that candidate K was elected and in terms of average vote share, candidate K performed better in the treatment than in the control condition in a robust manner. In particular, the biased feedback mechanism increased the vote share of the favoured candidate K by an average of 20 percentage points, 11.7 percentage points and 7.3 percentage points in E1, E2 and E3, respectively. These differences are very consistent across sessions and rounds and their magnitudes are meaningful politically. Importantly, these effects do not go away as participants gain more experience.

There is some evidence in E1 that learning fails in our environment, allowing biased polls to distort democratic outcomes over prolonged periods. It seems that the self-confirming nature of biased polls limits the scope of receiving the type of feedback that would reveal the bias. Perhaps more remarkably, explicitly informing voters in E3 about the biased rule for revealing poll results does not eliminate the electoral advantage that the bias yields to the 'favoured' candidate K. This indicates that voter unawareness about

the bias of polls is not the only factor that drives our results. Even when they are aware, subjects do not appear to rationally weigh the information content of polls. Instead, it seems that, in forming their expectations about the electoral results, voters use polls merely as judgemental anchors, so they overweight the reference point that polls provide and they underweight their informational content. This interpretation is consistent with the well-known process of anchoring-and-adjustment (Tversky & Kahneman, 1974).

An analysis of the relationship between participants' beliefs (regarding the election winner) and the revealed poll results supports the aforementioned mechanisms. We find that these beliefs are highly correlated with the average vote shares in the revealed polls, both in the treatment and in the control setting, especially in E1 and E3. Econometric results further indicate that beliefs do not increasingly deviate from revealed poll results as time passes. This means that voters do not discard or discount poll results in later rounds, indicating that very limited learning takes place. Moreover, averaged revealed poll results are a good predictor of electoral results in the treatment condition for all three experiments, although these polls were selected in a biased manner.

Overall, there is only weak evidence that participants are either able to realise the biased nature of polls or that they can sufficiently account for it when they are informed of it. These failures lead subjects to overestimate the popularity of the favoured candidate and to vote for her more frequently. Therefore, biased polls can have a significant and robust impact on election outcomes even when the public knows or suspects the bias. Consequently, our experiments may inform the public debate on whether or not biased polls can skew behaviour in real election settings.

The rest of the paper is structured as follows. Section 4.2 places our findings in the relevant literature. Section 4.3 presents our observational study on the propagation of poll results on social media. Section 4.4 discusses the design of our three experiments. In Section 4.5 we present descriptive results of our experiments, whereas in Section 4.6 we conduct regression analysis. Section 4.7 presents a short discussion of our findings and concludes.

4.2 Related Literature

The effects of polls on election outcomes have been the topic of both theoretical and empirical study. This large literature contains important experimental studies, but as far as we can tell, none of them considers biased feedback on actual polls along with opportunities for learning. Economic experiments have examined a variety of mechanisms that can drive poll effects on elections, with neutral phrasing and a theory-testing focus. An important mechanism examined in the lab is asymmetric information among voters (McKelvey & Ordeshook, 1984, 1985; Brown & Zech, 1973; Sinclair & Plott, 2012). This experimental strand finds that polls aggregate information reasonably well, although voters exhibit some robust elements of bounded rationality. A second studied mechanism has been coordination and strategic voting in multi-candidate elections (Forsythe, Rietz, Myerson & Weber, 1996; Plott, 1982), where the evidence indicates

that polls can often be instrumental in coordinating voters' choices. An additional important mechanism is turnout under costly voting. Most studies (Klor & Winter, 2007; Agranov, Goeree, Romero & Yariv, 2017; Gerber, Hoffman, Morgan & Raymond, 2017) point to a failure of the standard prediction that polls discourage majority group voting and that they are welfare reducing (Goeree & Grosser, 2007), although the effects seem generally complex.

However, the economics literature is mainly focused on unbiased polls, whereas our paper is concerned with biased polls and their effects on voting behaviour.[4] This is closer to the approach taken in political science, where many experiments strategically manipulate the poll information that participants receive. Typically, these experiments are non-incentivised. The early study by Fleitas (1971) indicates that voting is not responsive to the quantitative information revealed in polls. Meffert & Gschwend (2011) present different versions of newspaper articles that report voter support for German parties in multicandidate elections, while Rothschild & Malhotra (2014) manipulate the ostensible public support for several important issues and examine how this affects subjects' stated preference on the issues. These studies find that manipulation affects beliefs and moderately alters behaviour. Gerber et al. (2017) conduct large field experiments where they selectively convey poll results to manipulate the ostensible closeness of the race. Again, beliefs seem to be affected by the manipulation but behaviour not so much. As with previous experiments, rational choice theories, which predict voter turnout, do not perform very well.

The main difference between the aforementioned political science studies and ours is that these studies are not examining whether subjects are capable of understanding that manipulation is taking place and of accounting for it. In particular, in these studies participants face biased or manipulated polls only once, so they do not learn from past mistakes. Our design allows for multiple rounds of repetition so that we explore the participants' scope for learning. We believe that this is a critical aspect, as a standard approach of rational choice theory is to consider 'equilibrium' behaviour, i.e. stable patterns of behaviour after the effects of learning have taken place. In addition, our experiments show that voters are influenced by biased polls even when they are aware that polls are biased, a test that is absent from the aforementioned papers. This indicates that biased polls influence voters through multiple channels. To the best of our knowledge, no other study has attempted to disentangle the factors driving the effects

[4] We suspect that at least part of the reason for this omission in the experimental economics literature is reluctance to use what can be viewed as explicit manipulation in the lab. For instance, we refer to several studies in political science that expose subjects to different poll results (sometimes fabricated) and examine how this affects their behaviour. In our experiments, we avoid this approach that would unambiguously qualify as deception and we only provide truthful information. Still, some colleagues would count as deception any omission of information, as long as subjects are expected to behave differently in the presence of this information. However, most experiments where information is a treatment variable can be considered problematic under this strict definition. Moreover, according to this very strict approach, even information about other subjects' behaviour, or about the research objectives, should be shared with all subjects, but of course this would sometimes jeopardise the research design. We argue that the question of whether and how people are able to identify biased information can and should be examined in the economics laboratory, and how subjects form beliefs about whether information is biased or not should be an open research question, not a forbidden one.

of biased polls on election outcomes. Finally, our study is conducted in a laboratory and decision-making is incentivised with real money.

4.3 Bias in Online Propagation of Poll Results: Evidence from the Field

Although we mentioned several plausible mechanisms that could result in selective exposure of the public to poll results, most of them are difficult to study empirically. For instance, we do not have access to the pool of all methodologies that pollsters have at their disposal, neither are there published data on the set of poll results available to the media when they choose what news to broadcast. For this reason, we shall resort to showing that biased exposure can also ensue from the natural structure of modern political communication, namely social media.

To what extent do online publics paint a representative portrait of the existing results of opinion polls? We shall show that the nature of social networks results in a biased exposure of the public to poll results. People have various cognitive mechanisms that result in selective attention, such as negativity bias (Soroka, 2014), motivated reasoning (Taber & Lodge, 2006), cognitive dissonance (Morwitz & Pluzinski, 1996) or disproportionate responsiveness to outliers. Users of social media are also not demographically or politically representative of the general population (Mellon & Prosser, 2017), which could give rise to further biases in attention, via selective reporting. As a result, individuals attend to and, crucially, propagate to others, the results published by polling firms in a systematically biased manner.

In this analysis, we examine the biased propagation of published opinion poll estimates or trackers in the United States (US) and the United Kingdom (UK). Specifically, we consider measures of voting intentions for US Presidential elections (reported by HuffPost pollster.com) and for UK parliamentary elections (YouGov's political tracker). This enables us to assess the spread patterns of the published poll results in two different countries. Our objective is to show that in some real-life electoral races a subset of voters is exposed to poll results in a manner that systematically depends on the results of the polls themselves.

4.3.1 Opinion Polling in the US and UK

While opinion pollsters in the US and UK ask a wide variety of survey questions on political issues, among the most prominent measures of political attitudes are for presidential elections (in the US) and Westminster voting intentions (in the UK). These are central to depictions of the 'horse race' by media (Iyengar, 1991; Matthews, Pickup & Cutler, 2012). In the US, George Gallup famously introduced random sampling methods to measure national voting intentions in the 1936 presidential election. Variants of the question "If the election were held today, whom would you vote for?" have been asked regularly ever since. During the 2016 presidential election campaign there were well over

400 national opinion polls of voting intentions for Donald Trump and Hilary Clinton, yielding a steady flow of information on the election horse race.

In the UK, pollsters have been asking people about their voting intentions for Westminster parliamentary elections since 1943 (Wlezien, Jennings, Fisher, Ford & Pickup, 2013; Sturgis et al., 2016). YouGov has become one of the highest volume pollsters in the UK since their introduction of online methods in 2001, regularly fielding the question "If there were a general election held tomorrow, which party would you vote for?" During the first government of the UK's former Prime Minister David Cameron (2010-2015), it fielded a survey almost every other day.

4.3.2 Opinion Polling Data on Social Media (Twitter)

HuffPost Pollster and YouGov each report their latest poll estimates via their official accounts on the social media platform Twitter (in the case of HuffPost this involves polls conducted by other polling firms). This provides a regular stream of poll information that enables us to analyse patterns of selective reporting, by social media users, in an observational setting. With frequent estimations of public opinion (at least every other day), most fluctuations in poll estimates are attributable to noise due to sampling error (even where dampened by poll aggregators), and thus most users are (arguably) reacting to random short-term fluctuations, rather than systematic trends.[5] While it would in theory be possible to collect data on wider engagement with poll estimates on Twitter, this approach enables us to model a fairly stable source of poll information.

We obtained relevant tweets of poll estimates from @pollsterpolls and @YouGov using an advanced Twitter search with terms corresponding to the standard form of poll reporting used by each organisation (removing all extraneous cases from the scraped data). Details of these search terms are provided in Table 4.1. All the tweets report the current *level* of voting intentions for the relevant candidate or party. We calculate the change in voting intentions from the previous poll estimate in our dataset. This forms the independent variable of our analysis – the change in observed poll estimates.

[5] We will show that a systematic bias in the propagation of poll results is even evident in responses to short-term noise, rather than more sizable long-term trends. With such trends we should expect such bias to play an even more important role.

Table 4.1: Twitter reporting of poll estimates in the US and the UK

	US – Presidential election voting intentions	UK – General election voting intentions
Choice	Trump/Clinton	Labour/Conservatives/Liberal Democrats
Pollster	All pollsters	YouGov
Start	8 September 2015	9 April 2010
End	8 November 2016	8 December 2017
Measure	Voting intention, by candidate	Voting intention, by party
N of polls	445	1,451
N of days	428	2,801
Polls per day	1.04	0.52
N of Δ in vote	444	1,450
Retweets	5,054	41,291
Source	@pollsterpolls	@YouGov
Search terms	"2016 General Election", "Trump", "Clinton"	"Lab", "Con", "Westminster voting intentions"

Crucially, we also collected data on the number of 'retweets' for each tweet. This provides us with a measure of online propagation of the poll result, our dependent variable. On average, each poll estimate received 24.4 retweets, with an upward trend over time in the number of retweets as usage of Twitter grew. Our analysis undertakes an ordinary least squares regression of the number of retweets of a given poll estimate (*Retweets*) as a function of change in candidate or party support ($\Delta\, Vote$). In the US we focus on change in the 'margin' between the candidates, i.e. the lead of Clinton over Trump. In the UK we focus on change in support for the Labour, Conservative and Liberal Democrat parties. This focus on *change* enables us to determine whether biased propagation of poll results can stem from mere short-term fluctuations, rather than structural differences between particular candidates or parties.[6] The estimated models therefore take the following form, where *Equation 4.1* refers to the US, and *Equation 4.2* to the UK.

$$US: Retweets = a_0 + b_1\Delta(Vote(Clinton) - Vote(Trump)) + \epsilon \tag{4.1}$$

$$UK: Retweets = a_0 + b_1\Delta Vote(Con) + b_2\Delta Vote(Lab) + b_3\Delta Vote(LD) + \epsilon \tag{4.2}$$

The results for this analysis are reported in Tables 4.2 and 4.3. These reveal largely consistent, and also interesting, patterns (both within and across countries) in the propagation of poll estimates. In the US (Table 4.2), an one-unit increase in the Clinton-Trump lead in polls reported by HuffPost Pollster was associated with 1.0 additional retweets of the poll estimate. This might signify the partisan lean of the users of Twitter or the followers of this specific polling account, but it does hint at a selective reporting mechanism of poll estimates that fundamentally distorts voters' (salient) information on

[6] This focus on short-term dynamics enables us to show that even between adjacent days there is a systematic bias in propagation, and in particular selective propagation occurs regardless of how popular a candidate or party is.

the popularity of candidates. We should emphasise that we wish to establish empirically the existence of this distortion, and we do not wish to claim causality.

Table 4.2: Selective propagation of poll estimates of the US 2016 presidential election

	Retweets
Δ (Clinton-Trump)	1.021
	(0.148)***
Intercept	11.353
	(0.666)***
N	444
R-squared	0.10
Adjusted R-squared	0.10

$*p < 0.05; **p < 0.01; ***p < 0.001$

In the UK (Table 4.3), we see a similar pattern whereby a one-unit increase in voting intentions for the Conservative Party leads to 2.5 fewer retweets of the poll. In contrast, a one-unit increase in support for Labour leads to extra 7.3 retweets. There are no systematic differences for the Liberal Democrats, at least during this period. Predicted values of the regression models are depicted in Figure 4.1. These confirm the findings: there are distinct partisan differences in the online promulgation of poll results, specifically a bias where increases in support for left-wing parties/candidates are propagated more in online platforms, whereas it is drops in that support that receive wider spread for right-wing parties/candidates. In the UK context, interestingly, the pattern is more pronounced for Labour than the Conservatives, so this is not a purely symmetrical relationship.

Table 4.3: Selective propagation of poll estimates of voting intention, UK

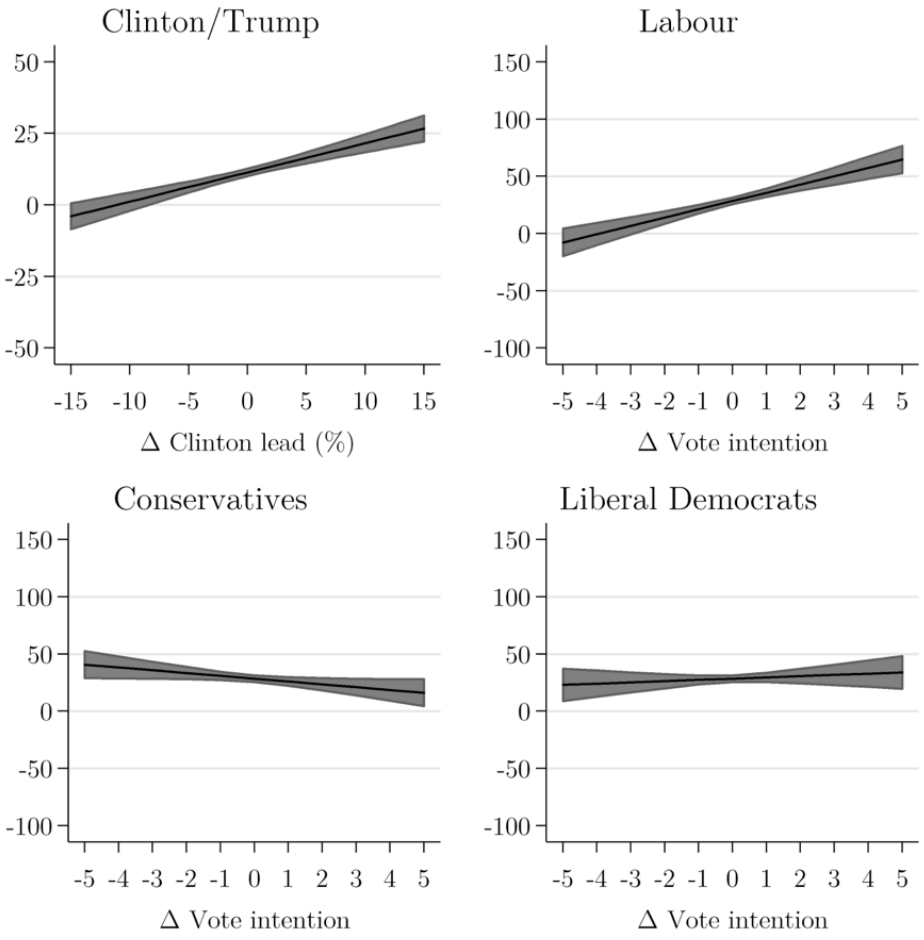
	Retweets
ΔVote(Con)	-2.464
	(1.162)*
ΔVote(Lab)	7.335
	(1.187)***
ΔVote(LD)	1.110
	(1.419)
Intercept	28.423
	(1.480)***
N	1,450
R-squared	0.04
Adjusted R-squared	0.04

$*p < 0.05; **p < 0.01; ***p < 0.001$

We have thus shown empirically that in modern democracies the public is likely to be exposed to the results of pre-election polls in a biased manner.[7] Now we may ask: what are the implications of such a bias for democratic elections? If biased exposure skews elections, then we should be concerned and maybe need to address this by policy changes. The problem is that establishing a causal relationship about such a complex phenomenon in the field can be very difficult. For this reason, we employ the experimental method increasingly popular in economics and political science (Palfrey, 2016). The virtue of this approach is that it can establish causality and discover general patterns of social behaviour in a controlled setting.

[7] Of course, given the network structure of the social media, it is not true that the same biased sample of polls is revealed to every voter (as is the case in our experiment). For instance, if left-leaning people mostly re-tweet to other left-leaning people, right-leaning people may not be exposed much to those polls. In fact, there is evidence that "the network of political retweets exhibits a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users." (Conover, Ratkiewicz, Francisco, Gonçalves, Menczer & Flammini, 2011). Accordingly, we are not claiming that the empirical pattern established here always generates a biased propagation pattern similar to the one used in our experiments. However, our correlational results do establish that different subsets of voters are exposed more to a certain type of results rather than to another type of results. As stated before, we believe that there are other propagation mechanisms that could more plausibly lead to an information pattern similar to the one used in our experimental design.

Figure 4.1: Adjusted predictions (with 95% confidence intervals) of the number of retweets, by ΔVote



4.4 Our Experimental Environment

In general, the information conveyed by poll results can be relevant to voters for many reasons (e.g. voting is costly and voters need to estimate the closeness of the race, there are multiple candidates and voters need to focus on a viable candidate, voters have bandwagon preferences, etc.). The particular environment we choose to study here is akin to Feddersen & Pesendorfer (1997), where voters assess candidates on two dimensions, their ideological position and their intrinsic quality (valence). In our setting, there are two political parties, party K and party J, each one of which fields a candidate. We refer to the candidates' identity by the name of the political party they stand for, hence the candidates are K and J.

All voters know the closeness of the candidates' political views to their own, i.e. the ideological position of the two candidates, but they differ in their knowledge of the candidates' valence. Some voters are informed and know precisely the valence of each candidate, while the remaining are uninformed and they know only the statistical

distribution out of which each valence is drawn. Moreover, in our setting informed voters are on average left-wing leaning in terms of ideological positions, while uninformed voters are on average right-wing leaning, so the voting intentions of the informed voters are not representative of the overall population. As a result, elections across the entire set of voters (not within the set of informed voters only) are meaningful for the aggregation of the electorate's preferences, while pre-election polls convey valuable information to uninformed voters by helping them make inferences about candidates' valence. In our setting, we have five voting intention polls taking place prior to each election.

Our research question, then, focuses on whether election outcomes are *affected* by giving voters a biased sample of the total information (total information in every round consists of the results from five polls), which systematically depicts the candidate of party K performing 'better' than in reality. This 'biased selection' environment constitutes our experimental *treatment manipulation*. We define the concept of 'affected' italicised above relative to two control conditions as benchmarks. Our first control (in E1) is simply an environment where the total information is released to voters. Our second control (in E2) is a setting where an equal amount of information as in our treatment manipulation (two polls out of five) is conveyed, but in a random, rather than a systematically selective, manner. The second control allows us to test whether the difference between observing all five polls and two selected polls is due to disparate quantities of information, i.e. observing a smaller set of polls (two instead of five), or whether it is due to the selection per se.

In a final experiment (E3), we also test whether the effect of biased polls is due to subjects perceiving the polls as unbiased (despite the feedback that they receive in every round) or due to their inability of properly inferring from feedback which is systematically biased (and subjects know this fact). We perform this test by replicating E1 with one important modification. In particular, in the treatment condition, participants are informed explicitly about the (biased) selection rule. All experiments are described in detail in Table 4.4.

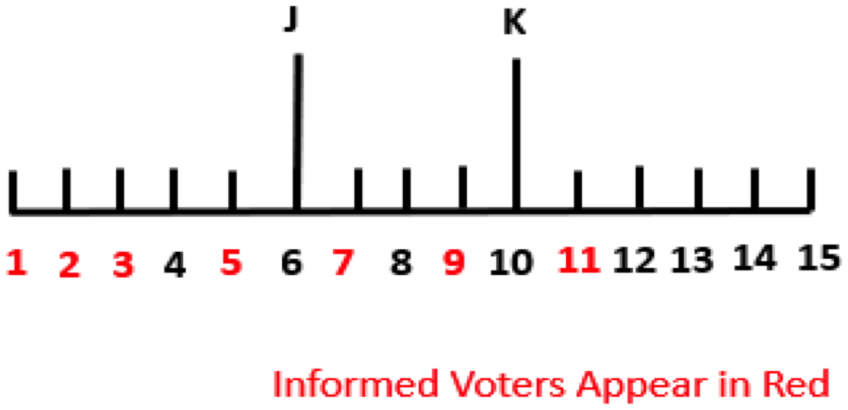
4.4.1 Voters' Preferences on Candidates, Voter Information, Polls, and Elections

In each experimental session, there are fifteen human voters (the two non-human 'candidates' are inactive, hence they do not vote). Voters are ordered according to their ideological positions as illustrated in Figure 4.2. Voter 1 is the most left-wing voter, while Voter 15 is the most right-wing voter. The median voter is in 'position 8', while candidates of parties J and K are in 'position 6' and in 'position 10', respectively. Ideological positions of candidates are the same in all rounds[8] and all voters know it in advance. At the beginning of each round, the ideological position of each voter is randomly drawn from integers between 1 to 15 (inclusive) without replacement.

[8]The interpretation is that the two parties consistently pick candidates that share their ideological views.

Each candidate's valence is drawn at the start of every round from a uniform distribution with values between 0 and 120.[9] At the time of the polls and the elections, the two drawn valences are known to voters in ideological positions $\{1, 2, 3, 5, 7, 9, 11\}$ who are the *informed voters*. The remaining voters, i.e. the ones in ideological positions $\{4, 6, 8, 10, 12, 13, 14, 15\}$, are the *uninformed voters*. They only know the distribution out of which the quality (valence) of the candidates is drawn.

Figure 4.2: Ideological preferences in the experimental interaction



The utility that voter $i \in \{1, 2, ..., 15\}$ obtains in the case where candidate $h \in \{J, K\}$ wins the election is given by $U_{ih} = X_i - \alpha d_{ih} + Q_h$, where U_{ih} is voter i's overall utility from candidate h being elected, X_i is voter i's utility from having a candidate with the same ideological position as herself being elected, while d_{ih} is the distance between the ideological positions of voter i and candidate h. Q_h is the valence of candidate h, and α is a parameter that measures the utility loss per unit of distance in ideological positions between i and h. For the purposes of our experiments, we set $X_i = 100$ and $\alpha = 5$ (for all voters, rounds, and sessions) and, as stated previously, $Q_h \sim U[0, 120]$.

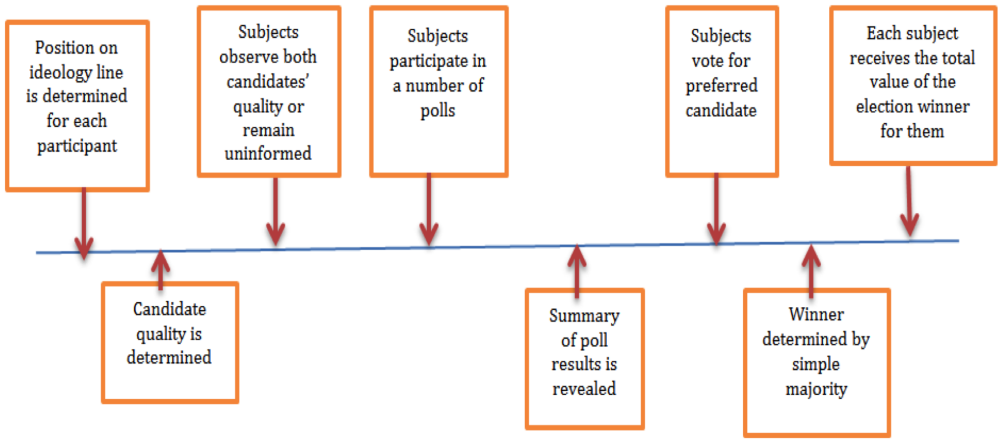
Since some voters are uninformed about the difference in valence between the two candidates, pre-election polls can be socially valuable in this setting. In particular, they can be utilised to transmit information about the candidates' valence from informed to uninformed voters. As explained earlier, it is important that the distribution of informed voters is not symmetric in the ideological spectrum. If that were the case, then the socially efficient outcome would be for uninformed voters to abstain from elections and let voting amongst informed voters determine the election outcome. In such an environment, polls would not perform a politically valuable role because participation of uninformed voters would not be necessary. Instead, polls are meaningful in our setting, because they aggregate information about candidate valence when the ideological preferences of informed voters do not represent the ideological preferences of uninformed voters.

[9] To reduce noise across sessions, we drew these valences once and for all before the start of the first session and used the same random draws for every session and for all experiments.

For example, assume that some uninformed voter observes substantial support in favour of K in the polls. If she perceives polls as unbiased and other subjects as rational, she will infer that K's valence is higher than J's, since some informed voters who are close to J's ideological position prefer to vote for K. These voters would do so only if K is of significantly higher valence than J. Accordingly, the uninformed voter, who observes the polls herself, infers from them the higher valence of candidate K and she may herself change her voting intention from J to K, if she is sufficiently close to the centre-left of the political spectrum.

After the valence is drawn for both J and K and informed voters receive this information, five polls, each inquiring four randomly chosen voters, take place. Sampled subjects are asked for whom they would like to vote in the upcoming elections (they may choose not to participate).[10] Given the number of drawn subjects who choose to participate, a poll reports the fraction of those in favour of K and in favour of J, respectively. For example, a poll revealing the following fractions: [25% for J, 75% for K] indicates that out of the four voters, all of whom chose to answer, three expressed support in favour of candidate K and one in favour of J.[11] Note that a single voter may participate in multiple polls. After the five polls are created by the above process, some subset of the results (depending on the experimental condition) is presented to all voters. The summary of each experimental round (as it was provided to subjects) is illustrated in Figure 4.3. The winner of the election is determined by simple majority, with ties broken by a 50-50 coin toss.

Figure 4.3: Sequence of actions in each experimental round



[10] Voters choose from the following three options: 'K', 'J', and 'Prefer not to participate'. Polls do not contain information on 'non-participation', as this substantially simplifies the feedback that subjects observe about the results of polls. This is especially important, since subjects need to infer overall support for each candidate on the basis of results from multiple polls.

[11] If, out of the four sampled voters, three opted to support K and one chose not to participate, the poll would be presented as 0% in favour of J and 100% in favour of K.

4.4.2 The Three Experiments

The only stage (out of those displayed in Figure 4.3) that differs across the two experimental conditions in each of our three experiments is the one where "summary of poll results is revealed". Table 4.4 describes our experimental design and Table 4.5 illustrates the information revealed in the two experimental conditions. Finding meaningful differences between 'control' and 'treatment' would indicate that biased polls can skew elections. The first benchmark (the control condition in our first experiment), which we use to judge whether 'skewing' takes place, is a perfectly transparent regime where all existing information (i.e. all five polls) is available to the public. This is a natural starting point. We also consider another benchmark (the control condition in our second experiment) where two out of the five polls are revealed in a random manner.

In terms of the treatment conditions, our natural point of departure (in E1 and E2) is an environment where voters observe the revealed information and have no a priori knowledge concerning how the two polls out of five are chosen to be revealed. In our view, this corresponds to many natural election environments of interest, where voters are not provided with any 'manual' describing the possible biases or agendas of those that reveal poll information. Instead, they have the chance to infer such biases and agendas through experience. In our experimental setting, this is accomplished because voters can compare poll predictions with actual election results (which they observe at the end of every round in all of our experimental conditions). In Experiment E3, we examine the consequences of providing a priori information about the exact nature of the bias to voters. Table 4.4 summarises the three experiments and the relevant 'control' and 'treatment' conditions in each one of them.

E1 and E2 had 120 participants each,[12] with eight 15-subject sessions (four control sessions and four treatment sessions).[13] E3 had 135 participants, with four control sessions and five treatment sessions. Participants in E1 and E2 were students at the University of Southampton and Newcastle Business School, and the experiments took place between May and November 2018. Participants in E3 were students at the University of York, and the experiment took place in June 2019. Our objective was for each experimental block (of 30 subjects) to achieve perfect randomisation by containing one control and one treatment session, with participants being randomly allocated between the two.[14]

In each session, subjects read instructions from their computer screens.[15] After the

[12] We shall use the words 'session' to denote each experimental interaction among 15 subjects who vote in the same 18 rounds (three practice rounds and 15 real rounds) of elections, and 'block' to denote the two sessions (one control and one treatment) taking place at the same time in the lab. A block has 30 subjects.

[13] We denote individual sessions as Ei_Cj or Ei_Tj where $i \in \{1,2,3\}$ denotes experiment, $j \in \{1,2,3,4,5\}$, denotes session , 'C' stands for control, and 'T' for treatment. For instance, $E1\_C1$ denotes the first control session in $E1$ and $E2\_T1$ the first treatment session in $E2$.

[14] The only three exceptions in this approach were sessions $E1\_C2$, $E1\_T2$ and $E3\_T5$, which were the only sessions of their block because of insufficient subject participation or lab capacity constraints.

[15] We programmed the experiments using O-tree (Chen et al., 2016) and recruited subjects via ORSEE (Greiner, 2015) in the University of Southampton and via *hroot* (Bock, Baetge & Nicklisch, 2014) in the Universities of Newcastle and York.

instructions, subjects participated in 18 rounds of play, including three practice rounds. At the end of the session, they were asked to complete a short questionnaire and were informed about their final score and monetary earnings. The core design of each round has been summarised in Figure 4.3. The only aspect that was not described is the 'belief elicitation' stage. In particular, after the release of the polls, participants were asked to state their beliefs about the vote shares of the two candidates in the elections. The information about polls took the form of a single probability distribution for each result, as shown in Table 4.5. Subjects' beliefs at the elicitation stage were also described in terms of this binary probability distribution.

Table 4.4: The experimental design

	E1	E2	E3
Treatment	The two polls (out of the five) with the greatest support for K are revealed.	The two polls (out of the five) with the greatest support for K are revealed.	The two polls (out of the five) with the greatest support for K are revealed. Subjects are a priori informed about this.
Control	All five polls are revealed.	Two out of the five polls are randomly revealed. Subjects are a priori informed about this.	All five polls are revealed.

Table 4.5: Example presentation of poll results in each condition

Treatment					
COMPANY		B			E
Candidate K		75%			100%
Candidate J		25%			0%
Control					
COMPANY	A	B	C	D	E
Candidate K	33%	75%	25%	67%	100%
Candidate J	67%	25%	75%	33%	0%

Notes. There are five polling companies, A to E. The result of each company is represented in terms of the two fractions measuring support for each candidate. In the control of E1 and E3, all five results are revealed, in a format similar to the example of the table. In addition, if the above table represented an actual set of poll results, then, in the treatment condition of all three experiments, companies B and E would be revealed, since these polls yield the highest support for candidate K.

4.5 Results and Descriptive Analysis

Let us first provide an overall summary of the *primary treatment effect* across the three experiments: the rate of electoral success. Table 4.6 illustrates the number of rounds won by each of the two parties in the treatment and control conditions across the three experiments. As can be seen, in E1 party K won 60% of all rounds in the control condition but 80% of the rounds in the treatment condition. In E2 party K won 61.6% of all rounds in the control but 73.3% of the rounds in the treatment, while in E3 party K won 56.7% of all rounds in the control but 64% of the rounds in the treatment. As we shall see in detail later, these differences are relatively homogeneous in their magnitude and extremely consistent in their sign, both across sessions of a given treatment and across rounds of a given session. In terms of statistical significance of the differences in individual experiments, the difference is significant in E1 (Fisher exact test with one-sided alternative) but not so in E2 and E3. Still, the differences are politically significant and very consistent, as we shall illustrate now.

Table 4.6: Number (percentage) of elections won for each party in each treatment and results of Fisher's exact test

	E1		E2		E3	
	Control	Treatment	Control	Treatment	Control	Treatment
K	36 (60%)	48 (80%)	37 (61.7%)	44 (73.3%)	34 (56.7%)	48 (64%)
J	24 (40%)	12 (20%)	23 (38.3%)	16 (26.7%)	26 (43.3%)	27 (36%)
p-value	0.0138		0.121		0.245	

Notes. The alternative is that the number of rounds won in the treatment condition is higher than in the control condition.

4.5.1 Experiment 1

Recall that in E1, the 15 participants in each control session voted every period after having been exposed to the results of all five polls, while in the treatment condition, the respective 15 participants were exposed to the two polls that had the greatest voting intention for the candidate of party K (but this was not explicitly stated). The most important general finding is summarised by descriptive analysis: the treatment did offer a considerable advantage to party K. Biased exposure to polls increased both the likelihood of party K winning the election and its vote share. Figures 4.4a and 4.4b juxtapose the fraction of election rounds won by K and vote shares for K in treatment vs. control sessions. It is clear that the electoral performance of K is consistently better in all treatment sessions relative to any control session. K won more rounds than J in both the treatment and the control condition. This is, however, to be expected since (by pure chance) in most rounds the randomly drawn valence for K was higher than the drawn valence for J. In fact, in 11 out of the 15 regular rounds K has higher valence than J, and in 9 of those the difference in favour of K is over 20 points.

Figure C.1 in the appendix indicates that there is enough heterogeneity in the findings of the five polls, so that revealing a biased selection of poll results is meaningful. For almost all rounds of the treatment sessions, the vote share of K differs substantially across polls, so selecting the ones with the highest share gives a non-representative image of the average vote share of K.

Furthermore, the difference in vote shares does not appear only at the average level, but also for each individual round. Figure 4.4c shows (for both the treatment and the control conditions) the vote share that candidate K received in each round (averaging across the four sessions of each treatment). The figure indicates that 'treatment' rounds have consistently higher vote shares for K than 'control' rounds. In fact, vote shares in 'treatment' are higher than vote shares in 'control' for all rounds. This is important because it does not seem to be the case that the difference vanishes in the last few rounds. Accordingly, these data are consistent with the interpretation that participants behave as if they perceive polls in the treatment as unbiased: they do not seem to be discounting them, even after several opportunities for learning. We shall now delve deeper into this important issue.

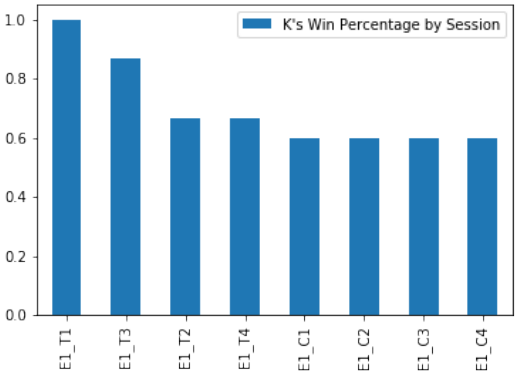
Evidence from Beliefs

As we explained previously, after the 'summary of polls' stage and before elections, participants were asked to state their beliefs about the vote shares of the two candidates in the upcoming elections. We use this elicitation of subjects' beliefs to examine whether they are in alignment with the poll information that participants received. If participants in the treatment condition perceived polls to be biased, then they should predict different vote shares for the election than the analogous poll information revealed, and this would lead to a low correlation between their beliefs and the average vote share in revealed polls. However, Figure 4.4d shows that the correlation is clearly not larger in the control sessions relative to the treatment sessions. Figures C.2 and C.3 in the appendix illustrate this relationship in more detail. In particular, they juxtapose (in each round and session) the average vote share of K according to revealed polls and the analogous vote share that subjects expect according to their average beliefs. In both conditions, average beliefs closely follow the average vote share revealed in polls, with no discernible pattern of differences. This is consistent with the idea that participants perceive polls as unbiased in both the treatment and the control condition.

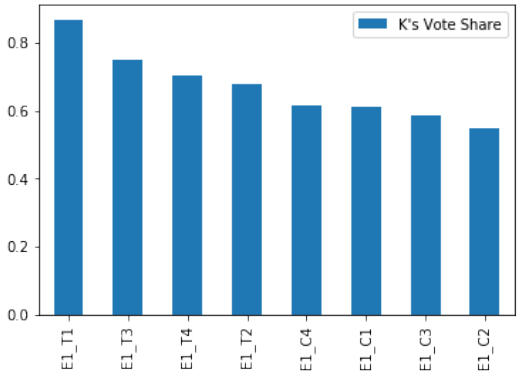
4.5.2 Experiment 2

E1 compared the electoral results in a 'biased regime', where there is a systematically biased selection of poll results revealed to the public, to a 'full information' regime. This full information regime is a natural benchmark to consider: the public is informed about the totality of relevant evidence for democratic decision-making. However, a weakness of this benchmark is that it provides more information than the control condition (the results of five polls instead of two). For this reason, it is important to also

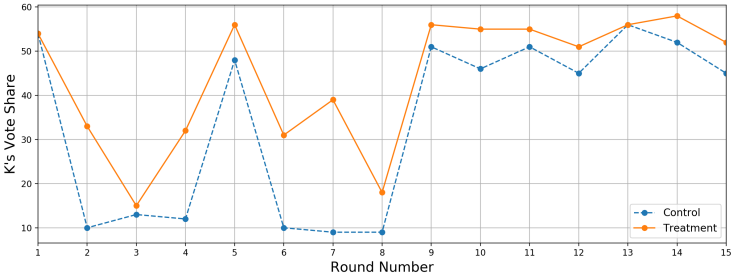
Figure 4.4: Descriptive results of E1



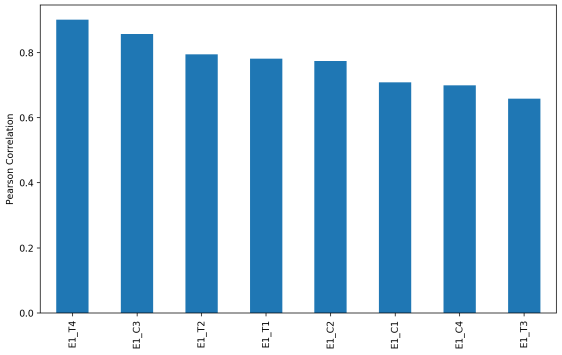
(a) Fraction of rounds won by K in each experimental session of E1



(b) Average vote share for K in each experimental session of E1



(c) Comparison of vote share round-per-round in E1



(d) Statistical correlation between elicited beliefs and poll results revealed in E1

employ a control condition where the amount of information is similar to the treatment (the 'biased regime'). For this purpose, we conducted the same number of randomised blocks (four 30-subject blocks) in an additional experiment (Experiment E2) where the control condition revealed the results of only two out of the five polls, and these two polls were chosen randomly.

Figures 4.5a and 4.5b contain the basic descriptive results from this experiment. The evidence points consistently to the direction observed in E1, but the treatment effects are smaller. This should not be surprising, if one considers the censored nature of the results revealed in the treatment condition. This entails that the treatment and control conditions are closer to each other in E2 (in terms of revealed poll results), than in E1. In other words, by pure chance the poll results revealed in the treatment and the control in E2 can be close to each other or even identical, which (almost certainly) cannot be the case in the comparison between treatment and control in E1.

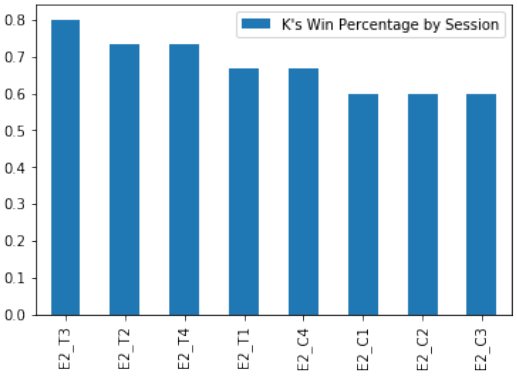
Figure 4.5c illustrates the vote share round-by-round. Once more, the pattern is that the vote share for K is uniformly higher in the treatment than in the control, while the difference does not seem to disappear with learning. These differences appear somewhat smaller than in E1. However, the difference is still meaningful: the number of rounds won by J in the control is nearly 50% larger than in the treatment (23 vs. 16). Overall, the consistency of the pattern indicates that the biased release of poll information has relatively robust and predictable effects on electoral results.[16]

Evidence from Beliefs

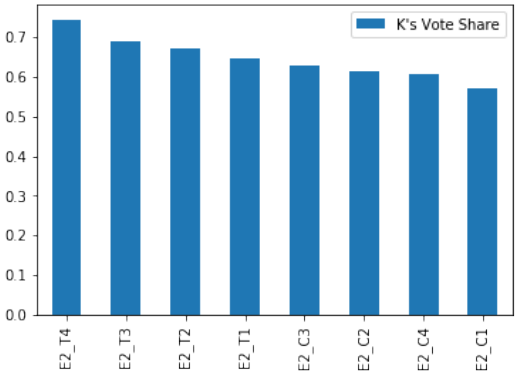
There is no a priori reason to expect that subjects in the treatment condition of E2 would behave differently than in the treatment condition of E1, since these conditions are practically identical. However, a careful inspection of Figures C.5 and C.6 of the appendix reveals that now some systematic patterns of discounting poll results might exist. In particular, in Figure C.6 it seems that average beliefs in the treatment sessions tend to be lower than average revealed poll results. This means that subjects seem to somewhat discount the (inflated because of bias) advantage in favour of K presented in the polls. This pattern does not seem to hold for the control sessions, as evidenced by Figure C.5. Figure 4.5d tends to confirm this pattern. In particular, it now appears possible that the correlation between average beliefs and revealed polls is systematically higher in the control (where information is unbiased) than in the treatment (where information is biased). However, as we shall see in Section 4.6, further econometric analysis cannot provide evidence for subjects discounting biased polls.

[16] For instance, we expect that if we were to conduct experiments with 10 polls instead of 5 (keeping other aspects of the experimental environment constant), we would likely find significant differences. However, implementing this would probably be too burdensome for participants in our current experimental environment.

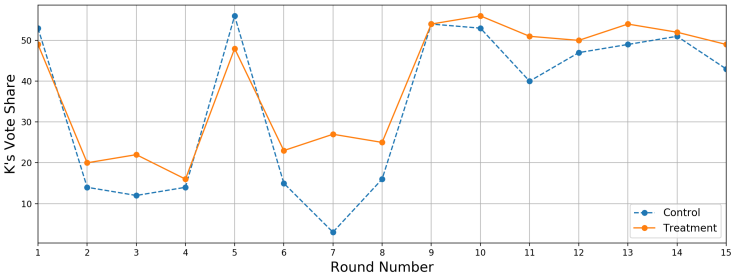
Figure 4.5: Descriptive results of E2



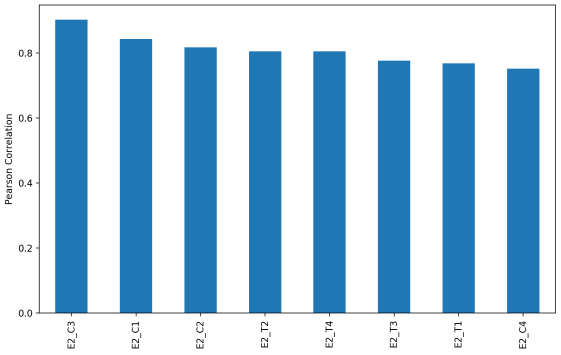
(a) Fraction of rounds won by K in each experimental session of E2



(b) Average vote share for K in each experimental session of E2



(c) Comparison of vote share round-per-round in E2



(d) Statistical correlation between elicited beliefs and poll results revealed in E2

4.5.3 Experiment 3

E1 showed in a particularly robust manner that election outcomes in a 'biased feedback' environment (where poll feedback is systematically selected to maximise the seeming popularity of a particular party) are distorted by this reporting bias. The effect is large in political terms: compared to the 'transparent democracy' benchmark, where the information from all polls is revealed, the 'biased feedback' environment resulted to an increase of the winning rate for party K (favoured by the biased feedback) by twenty percentage points. Further analysis will show that the experimental results in E1 are consistent with the notion that the systematically selected poll results become self-confirming, which prevents feedback that could have exposed the biased rule underlying them. Experiment E2 showed that the treatment effect is robust (but its magnitude is unsurprisingly smaller) when one accounts for the fact that the 'transparent democracy' benchmark has more information revealed than the 'biased feedback' condition.

However, it may be argued that in actual democratic elections people have enough experience with the political process and the media in order to gauge the agendas and incentives of those who reveal poll information. In particular, it is likely that most voters have a strong prior about the 'biased feedback' rule. Accordingly, our environment in the treatment conditions of E1 and E2 might be criticised as capturing only the special case of elections with young or inexperienced voters, especially in early rounds of play. Moreover, the structure of the treatment conditions of E1 and E2 make it difficult to pinpoint exactly the mechanism that drives the treatment effect. In particular, the effect may be either because of the inability of voters to understand that the information is selected in a systematically biased manner, or due to their difficulty in deducing information from a biased set of results even when they know the biased process that generates them.

To address these concerns, we run a third experiment (E3) where the treatment condition entails using the same biased rule as in the treatment conditions of E1 and E2, but with full clarity about this biased rule. In particular, the instructions mentioned that: "After polls have taken place in each round, the findings of the two companies which exhibit the greatest support for candidate K will be revealed to you. All participants will observe the fraction of votes that each of the two candidates received in the polls of these two companies" and then provided an example to illustrate the biased rule. In this environment, a rational participant would observe the results of these two companies and then try to gauge information about the valence of the two candidates accounting for the selection rule underlying these results. Once more, the issue is whether subjects sufficiently discount the information (typically) in favour of K having the higher valence, and thus whether society avoids the swaying of election results due to the biased reporting rule.

The basic results of E3 (which had five treatment sessions with the 'known biased rule' and four control sessions with the 'transparent democracy' information environment) are illustrated in Figures 4.6a -4.6c. As can be seen, even in this case, the biased feedback rule seems to offer an advantage to candidate K. In particular, the four sessions

with the best electoral performance for K (as measured by the fraction of elections won) are all sessions with the 'known biased rule'. The difference is – once more – politically meaningful: the number of rounds won by J per session in the control is about 20% larger than in the treatment (6.5 vs. 5.4). Again, it is the consistency and robustness of the effect of the biased release of poll information on electoral results that is striking.

A similar message is conveyed by examining the average vote share of K in each session. In particular, in all treatment sessions K has a higher vote share relative to any control session. Figure 4.6c additionally shows that the difference does not depend on the particular round, but that it is rather sizable for every individual round. Again, the picture that emerges is that biased exposure of the public to poll results affects elections in a pretty robust manner.

One interpretation of this finding is that polls create a judgemental anchor for voters' beliefs regarding election outcomes. Voters do not seem to have the capacity to account for the bias in the polls to its full extent. Instead, they seem to use poll results as anchors, which they adjust until they reach an acceptable range. The use of such as heuristic is reasonable, given the demanding learning required in this voting setting.

Evidence from Beliefs

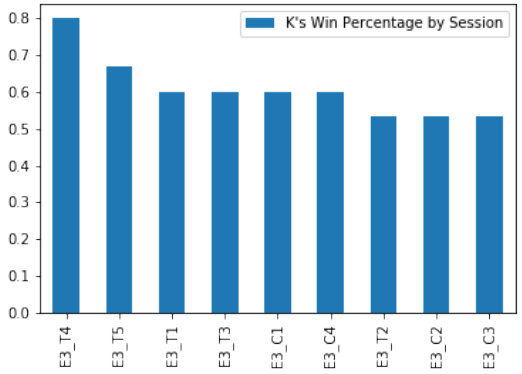
The comparison between average results of revealed polls and elicited beliefs becomes very interesting, especially compared to experiments E2 and E1. Figure 4.6d illustrates the correlation between average beliefs and revealed polls. Remarkably, no systematic pattern emerges. In particular, it does not appear to be the case that subjects account for the bias when they form their beliefs. Beliefs are not closer to revealed poll results in the unbiased control condition than in the biased treatment condition, a finding consistent with the belief correlations of E1, but not of E2. This is very surprising, since on the one hand there is a set of poll results known to be a biased sample from the available evidence on voters' preferences, and on the other hand there is the totality of the evidence. Yet, results indicate that subjects do not appear to distrust the first set of results more than the second one, reinforcing the interpretation that polls generate judgemental anchors for beliefs and adjustment is insufficient.[17]

4.5.4 Additional Descriptive Analysis

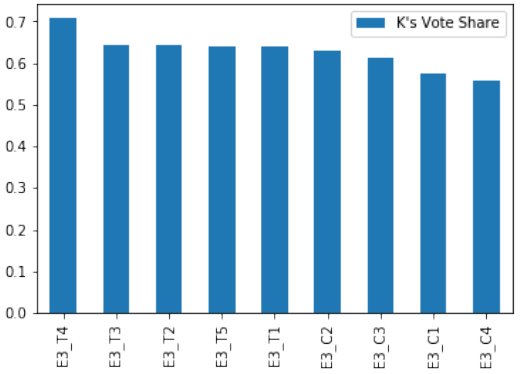
In terms of the welfare effects of biased polls, a rough measure of utilitarian welfare is the average experimental payoffs in each condition. Intuition suggests that the biased nature of polls should have a negative impact on this measure as the bias introduces noise in the information conveyed by polls to voters relative to the case of unbiased polls. Moreover, if voters do not discount the information of biased polls properly, then they

[17] In fact, there is some graphical evidence of some (very limited) adjustment. An inspection of Figures C.8 and C.9 in the appendix indicates that in the treatment sessions of E3 there might be a weak tendency for average revealed polls to exceed average beliefs. But neither the aforementioned correlation analysis nor our regression analysis corroborate the idea that meaningful discounting of poll results is taking place in E3.

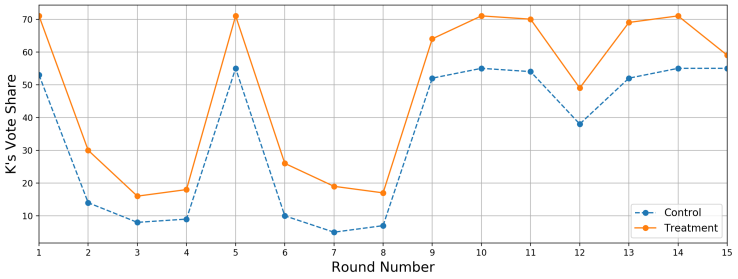
Figure 4.6: Descriptive results of E3



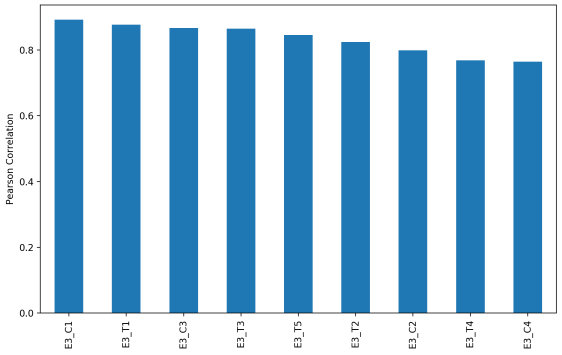
(a) Fraction of rounds won by K in each experimental session of E3



(b) Average vote share for K in each experimental session of E3



(c) Comparison of vote share round-per-round in E3



(d) Statistical correlation between elicited beliefs and poll results revealed in E3

will tend to vote more frequently for candidate K even if he is of lower valence than candidate J.

Indeed, our findings confirm these conjectures, as can be seen in Table 4.7. In particular, sessions in the treatment condition were generally associated with lower payoffs per subject than sessions in the control condition. In fact, average individual payoffs across conditions were 171.3 (control) vs. 165.5 (treatment) in E1, 169.8 vs. 167.4 (respectively) in E2 and 171.03 vs. 169.15 (respectively) in E3. This disparity resulted from the fact that the high-valence candidate lost in the treatment condition more often than in the control condition.

Specifically, in the control of E1 the high-valence candidate always won. In contrast, in the treatment condition of E1 there were 12 elections where candidate J lost, despite having the higher valence (the opposite direction was not observed). In E2, while in the control condition there were 3 elections where the high-valence candidate lost, this increased to 8 elections in the treatment condition.[18] In E3, in the control condition, out of 60 elections, there were two cases where K was the high-valence candidate but J won in the end. The opposite never happened. In the treatment condition, out of 75 elections, there were two times when K was the high-valence candidate but J won in the end, and four times when J was the high-valence candidate but K won in the end.

It is worth emphasizing that for an overall assessment of the social, economic and political implications of our experimental results, these utilitarian welfare effects are only of complementary importance. Changing the margins of victory also has important implications for the parliamentary representation of parties and for long-run political competition, and these aspects cannot be captured by this limited analysis.

It also worthwhile to provide some insights on the behaviour of informed voters. We should note that in our experiments, informed voters face an easy decision: they should simply vote for the candidate that gives them the higher payoff, which they can easily calculate.[19] Accordingly, if these individuals' votes deviate from 'optimal behaviour' this would indicate that the assumption of rational, money-maximising political agents is violated. Table 4.8 illustrates the behaviour of informed voters. For instance, in 8.57% of the 420 decisions that informed voters made in the control condition of E1, informed voters chose candidate K although the money-maximising choice was candidate J. Similarly, in 34.05% of the 420 decisions that informed voters made in the treatment condition of E2, informed voters chose candidate J and their money-maximising choice was also candidate J. As can be seen, most decisions by informed voters are consistent with the money-maximising model.

Nonetheless, a non-trivial fraction of decisions, slightly lower than 15% for the controls and ranging between 11% and 23% for the treatments, deviates from the prediction of the model of selfish money-maximising agents. A possible explanation for this behaviour is 'bandwagon preferences', i.e. a genuine willingness of the participants to vote

[18] Out of all these instances, only once did J win when K had the higher valence (it happened in the control condition).

[19] For simplicity, we shall call 'h-voter' an informed voter whose money-maximising choice is candidate h, where $h \in \{J, K\}$.

for the likely winner, which is not captured by monetary payoffs. Interestingly, J-voters are more likely to vote for candidate K in the treatment setting than in the control, and within the treatment setting this type of behaviour is more common than the opposite (i.e. K-voters voting for candidate J). Thus, 'bandwagon preferences' are likely to be relevant, and in particular they seem to amplify the effects of biased polls.

It is also important to discuss the behaviour of voters at the poll stage. Tables C.1 -C.3 in the appendix provide an overall summary of voting behaviour in the different sessions in our three experiments. The results are broken down by different status of voters (informed vs. uninformed). Certain insights can be inferred from Tables C.1 -C.3: informed voters are more likely to participate to polls, while uninformed voters are more likely to vote for K in polls (which makes sense, since their ideologies are closer to K). Moreover, there seems to exist no significant difference between treatment and control, which is again unsurprising, since the treatment is different from the control only when voters observe poll results.

Table 4.9 compares the voting choice at the poll stage to the one at the actual elections.[20] The table indicates that, if subjects truthfully reported voting intentions in the polls, the treatment induced some voters to switch in the direction of voting for K in the elections. Moreover, the voting pattern for those that chose K in the polls is similar across experiments: in all experiments, about 8-10% of poll voters for K, who would otherwise depart from voting K in the elections, are induced by the treatment to stick to K. However, there are significant differences across experiments in the behaviour of those who chose J at the poll stage, and these can partially account for the heterogeneity of the primary treatment effect across experiments. In particular, as we move from E1 to E2, and then to E3, the effects of treatment in inducing those that voted for J in polls to switch to K in the elections falls from 19.2% to 7.85% to about 1%.[21] These were mainly uninformed voters who are closer to J, but were induced to switch to K in the elections because of the treatment.

[20] Note that this table does not contain the behaviour of all subjects, since some were not randomly chosen to any poll, and some who were chosen opted not to participate. In total, Table 4.9 contains information for about 70% of overall decisions.

[21] These percentages are obtained as the difference between treatment and control in the J/K row in each experiment. Recall that the entries in this row correspond to the percentage of cases (out of all cases where a subject voted both in the polls and the elections) that a voter chose J in the polls but K in the elections. The higher occurrence of this in the treatment condition can be interpreted as a treatment effect.

Table 4.7: Average payoffs in each session

Session in E1	Average Experimental Payoffs	Session in E2	Average Experimental Payoffs	Session in E3	Average Experimental Payoffs
E1_C1	171.27	E2_C1	171.27	E3_C1	171.27
E1_C2	171.27	E2_C2	167.40	E3_C2	170.80
E1_C3	171.27	E2_C3	171.27	E3_C3	170.80
E1_C4	171.27	E2_C4	169.33	E3_C4	171.27
E1_T1	159.87	E2_T1	169.93	E3_T1	171.27
E1_T2	168.93	E2_T2	166.93	E3_T2	170.80
E1_T3	164.27	E2_T3	165.93	E3_T3	168.73
E1_T4	168.93	E2_T4	166.93	E3_T4	165.67
				E3_T5	169.27

Table 4.8: Behaviour of informed voters

Preferred/ Voted for	E1		E2		E3	
	Percent of total choices		Percent of total choices		Percent of total choices	
	Cont.	Treat.	Cont.	Treat.	Cont.	Treat.
J/J	38.33%	27.86%	37.86%	34.05%	38.57%	39.05%
J/K	8.57%	19.76%	8.81%	13.81%	8.81%	9.14%
K/K	44.52%	47.38%	46.43%	47.38%	46.38%	49.52%
K/J	5.24%	3.33%	4.29%	3.81%	3.10%	1.90%

Notes. 'Preferred' stands for the money-maximising choice of candidate, while 'voted for' signifies the actual voting choice in the elections. Please note that the fractions do not add up to 100%, because abstention is allowed at the election voting stage. In total, there are 420 decisions by the seven informed voters in the four sessions of each condition of each experiment (except E3, where in the treatment condition there are 525 such decisions).

Table 4.9: Comparison of individuals' voting at the polls vs. the final election

	E1		E2		E3	
poll/election	treatment	control	treatment	control	treatment	control
J/J	58.00%	76.06%	72.43%	78.35%	77.50%	78.69%
J/K	40.80%	22.01%	26.75%	18.90%	20.63%	19.67%
J/A	1.20%	1.93%	0.82%	2.76%	1.88%	1.64%
K/J	4.68%	13.99%	5.40%	12.57%	5.82%	14.04%
K/K	94.55%	84.55%	94.03%	86.03%	93.32%	85.67%
K/A	0.78%	1.46%	0.57%	1.40%	0.86%	0.29%

Notes. 'A' stands for abstention.

4.6 Regression Analysis

The preceding descriptive analysis shows that exposure to biased polls increases the likelihood of 'favoured' candidate K being elected. A key question is why this is happening. Because of the relatively complex environment we are studying, it is unlikely that voters use the strategic structure of the environment to predict behaviour deductively, thus we shall focus our analysis on the effects of feedback and learning on beliefs and behaviour. In particular, in E1 and E2, do subjects manage to learn that in the treatment condition the revealed polls are not a representative image of subjects' preferences at that particular time? What is the relationship between the poll information that subjects observe, their beliefs and election results? In the following, we shall employ our measures of subjects' beliefs and try to tackle these questions. The first model we estimate (Model 1) takes the following form:

$$B_t = a + b_1 P_t + b_2 R + b_3 (R * P_t) + b_4 T + b_5 (T * P_t) + e_t \tag{4.3}$$

We consider one round as the unit of observation, so data are at the session level. The dependent variable B_t is the average subjects' beliefs about candidate K's vote share in period t. P_t is the share of voters supporting K that can be inferred by the revealed polls in round t. For instance, in E1, in the treatment condition, this share is derived as the average of two polls, while in the control condition, this share is derived as the average of five polls.[22] R is the 'late rounds' dummy variable taking the value 0 for early rounds (rounds 1 to 10) and 1 for late rounds (rounds 11 to 15). T is the treatment dummy (1 if the session is in the treatment condition, 0 otherwise).

We are principally interested in the coefficients of the two interaction terms. A significant negative coefficient in the interaction term $P_t \cdot R$ would indicate that the degree to which revealed poll results affect beliefs weakens through time. This would be consistent with the notion that subjects distrust polls at the treatment condition (but we

[22] Thus, this specification models voters as rather unsophisticated, forming inferences about each candidate's support by merely taking the average of the polls revealed to them.

would not expect the same for the control condition). On the other hand, a significant negative interaction between P_t and T would imply that in the treatment condition there is a weaker relationship between beliefs and average announced poll results. Of course, we would expect such a negative interaction to exist in E3, since participants are explicitly informed about the bias.

As Table 4.10 indicates, the results of the model do not support the notion that subjects in E1 are able to learn and account for the bias in the treatment condition. In particular, there is no significant interaction between the 'late rounds' dummy and average announced poll information, although the respective coefficients are negative in both the treatment and the control. Table 4.10 show a similar pattern. Interestingly, the estimated coefficient b_3 is positive in the control but negative in the treatment setting in both E2 and E3. However, none of this is statistically significant. On aggregate, there seems to be very weak, if any at all, evidence that subjects somewhat discount poll information in late rounds. The experimental condition also does not seem to make a difference: the estimated coefficient b_5 does not have a consistent sign across the three experiments and it is not statistically significant in any of them.

Finally, with regards to the variable P_t, the results of Table 4.10 indicate that even if the bias is known a priori, there is a very strong relationship between beliefs and average revealed poll results. The estimated coefficient b_1 is positive and statistically significant at the 1% level in all settings and for all experiments. As expected, the relationship appears stronger in the control condition (although the difference is not statistically significant).

Table 4.10: Effect on Beliefs (Model 1)

	E1			E2			E3		
	Pooled	Treatment	Control	Pooled	Treatment	Control	Pooled	Treatment	Control
Avg. poll info.	0.858***	0.848***	0.856***	0.798***	0.787***	0.815***	0.887***	0.847***	0.897***
	(-0.029)	(-0.035)	(-0.034)	(-0.021)	(-0.029)	(-0.023)	(-0.025)	(-0.025)	(-0.032)
Late rounds dummy	12.350**	3.823	16.819*	-4.151	14	-2.863	3.268	7.117	-0.342
	(-6.183)	(-9.519)	(-9.718)	(-4.133)	(-12.397)	(-4.684)	(-4.573)	(-5.523)	(-8.055)
Late rounds dummy *Avg. poll info.	-0.114	-0.021	-0.166	0.08	-0.092	0.036	0.018	-0.015	0.044
	(-0.076)	(-0.109)	(-0.133)	(-0.049)	(-0.133)	(-0.062)	(-0.054)	(-0.062)	(-0.1)
Is treatment*Avg. poll info.	-0.005			0.002			-0.038		
	(-0.045)			(-0.033)			(-0.033)		
Is treatment	-0.74			-3.657			-2.041		
	(-3.245)			(-2.5)			(-2.442)		
Constant	9.185***	8.975***	9.094***	11.987***	8.667***	11.694***	5.825***	3.592*	5.750***
	(-1.711)	(-2.655)	(-1.895)	(-1.341)	(-2.15)	(-1.364)	(-1.605)	(-1.811)	(-1.823)
Observations	120	60	60	120	60	60	135	75	60
R-squared	0.945	0.931	0.942	0.963	0.954	0.969	0.96	0.962	0.958

Standard errors in parentheses.

$*p < 0.1, **p < 0.05, ***p < 0.01$

The second model we examine (Model 2) takes the form:

$$V_t = a + b_1 P_t + b_2 T + b_3 (T * P_t) + e_t \tag{4.4}$$

V_t is the vote share which K received in the election of round t. Election results are regressed on average poll information, the treatment dummy and the interaction term. This specification will allow us to examine whether average revealed poll results are good predictors of the electoral performance of K. In particular, if voters realise the bias and discount for it, we would expect that in the treatment condition polls have less influence on the final election. The results are presented in Table 4.11 For all experiments, we find that the coefficient of the interaction term is small and not significant. Moreover, the estimated coefficient of P_t is often larger in the treatment condition. There is clearly no evidence that revealed poll results are better predictors of election results in the control, rather than in the treatment condition.

The above finding is at face value paradoxical, since unbiased polls should be closer to election results than biased ones. However, this conclusion ignores the fact that revealed poll results might affect behaviour. The data indicate that the change in behaviour induced by polls can sometimes render the predictions of the biased sample of polls self-confirming – or, at least, nearly as good a predictor of elections as the unbiased sample. Figures 4.7a -4.7c illustrate the distribution of the differences between the average revealed poll results and the actual election results of the same round (both of these results are represented by the voting share for K). These figures can help us assess this issue further.

A visual inspection of Figure 4.7a indicates that, in E1, the average vote share of K according to the revealed poll results deviates from the election vote share for K no more in the treatment sessions than in the control sessions. If anything, the variance in the deviations seems to be larger in the control. On the other hand, Figures 4.7b and 4.7c indicate that, in E2 and E3, there is a pattern whereby in the treatment – but not in the control – average revealed polls systematically over-predict K's vote share.

This is important, as it might cast some light on the inability of participants (especially in E1 and E3) to account for the biased nature of poll results. Subjects who do not account for strategic incentives, but learn from experience alone (in the spirit of reinforcement learning) will have the opportunity to observe differences such as those presented in Figures 4.7a -4.7c for a number of rounds. If these observed differences are not systematically greater in the treatment condition than in the control condition, we should not expect that such naïve learners would discount the 'biased' revealed poll results of the treatment any more than the 'unbiased' results of the control. This is consistent with what happens in E1.

However, as we noted regarding E2, Figure 4.7b illustrates systematically larger disparities between revealed polls and election results in the treatment condition vs. the control condition. There is evidence that this resulted in subjects discounting somewhat the average revealed poll results in the treatment condition of E2: correlations

between these poll results and elicited beliefs are lower in the treatment condition (see Figure 4.5d). However, in E3, despite the fact that subjects are a priori informed of the bias in addition to the opportunity of observing a systematic overprediction of K's vote share in the treatment (see Figure 4.7c), there is not much evidence of discounting average poll results in the treatment condition.

Table 4.11: Effect on Average Poll Information (Model 2)

	E1			E2			E3	
	Pooled	Treatment	Control	Pooled	Treatment	Control	Pooled	Treatment	Control
Avg. poll info.	1.165***	1.182***	1.165***	1.018***	1.134***	1.018***	1.184***	1.139***	1.184***
	(-0.062)	(-0.066)	(-0.07)	(-0.052)	(-0.064)	(-0.055)	(-0.052)	(-0.05)	(-0.052)
Is treatment	-9.609			-19.668***			-8.06		
	(-7.287)			(-6.524)			(-5.319)		
Is treatment*Avg. poll info.	0.017			0.116			-0.045		
	(-0.099)			(-0.086)			(-0.072)		
Constant	-6.297	-15.907***	-6.297	0.334	-19.334***	0.334	-11.461***	-19.521***	-11.461***
	(-3.866)	(-5.268)	(-4.362)	(-3.446)	(-5.181)	(-3.656)	(-3.483)	(-4)	(-3.505)
Observations	120	60	60	120	60	60	135	75	60
R-squared	0.844	0.846	0.827	0.854	0.845	0.856	0.888	0.876	0.899

Standard errors in parentheses.

$*p < .1, **p < .05, ***p < .01$

The third model we examine (Model 3) considers differences:

$$\Delta BP_t = a + b_1 \Delta PV_{t-1} + b_2 T + b_3 (T * \Delta PV_{t-1}) + e_t \tag{4.5}$$

ΔBP_t equals B_t minus P_t and ΔPV_t is P_t minus V_t. We use Model 3 to explicitly examine whether there is evidence for learning. Again, we focus on reinforcement-type learners, who observe the model's variables through time. If they observe that ΔPV_{t-1} is large, this means that (in the last period) polls overestimated the performance of K relative to the election outcome. We expect that if subjects learn, this will result in subjects adjusting their beliefs (for K's share) downwards conditional on the poll results, hence we expect a decrease in ΔBP_t. However, as Table 4.12 indicate, the coefficients for ΔPV_{t-1} are small and not significant. Once more, we find little evidence that subjects, in forming their beliefs, adjust for the existence of biased polls in our experimental environment.

Table 4.12: Effect on the Differences between Beliefs and Average Poll Information (Model 3)

	E1			E2			E3	
	Pooled	Treatment	Control	Pooled	Treatment	Control	Pooled	Treatment	Control
ΔPV_{t-1}	0.034	-0.086	0.034	0.051	-0.032	0.051	-0.031	-0.064	-0.031
	(-0.058)	(-0.069)	(-0.062)	(-0.066)	(-0.067)	(-0.072)	(-0.065)	(-0.06)	(-0.063)
Is treatment	-4.662***			-6.826***			-6.042***		
	(-1.238)			(-1.421)			(-1.28)		
Is treatment *ΔPV_{t-1}	-0.12			-0.083			-0.033		
	(-0.094)			(-0.099)			(-0.087)		
Constant	3.083***	-1.579*	3.083***	1.304	-5.523***	1.304	1.118	-4.924***	1.118
	(-0.87)	(-0.819)	(-0.927)	(-0.893)	(-0.996)	(-0.974)	(-0.866)	(-0.963)	(-0.842)
Observations	112	56	56	112	56	56	126	70	56
R-squared	0.137	0.029	0.005	0.229	0.004	0.009	0.217	0.016	0.004

Standard errors in parentheses.
*$p < .1$, * * $p < .05$, * * *$p < .01$

4.7 Discussion and Conclusions

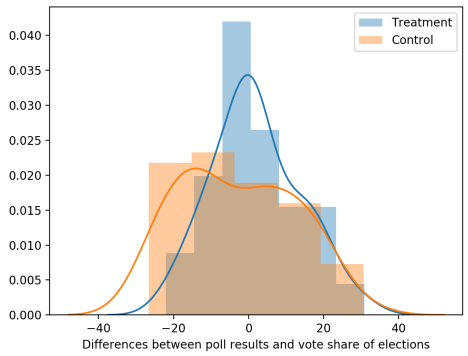
In this paper we examined the existence and implications of biased mechanisms that propagate the results of voting intention polls. We first established the existence of a systematically biased propagation pattern in the field, by analysing how news regarding the electoral 'horse race' are reproduced in social networks. The data indicate that 'good news' have a higher chance of being propagated in the examined network if they concern liberal politicians than if they concerned conservatives. We then presented results from a series of experiments with majority voting where participants received information regarding poll results in a systematically selective manner. The environment we considered is a two-candidate election contest with common values (concerning candidates' valence) and no voting costs.

Our findings indicate that exposure to biased polls consistently skews the electoral outcome in a predictable way. In a very robust manner, elections that took place in the 'biased polls' environment provided an electoral advantage to the party that was 'favoured by the bias'. This effect was smaller when in the control condition two polls were randomly revealed, as opposed to all five polls being revealed, but the direction of the effect was consistently the same. Similarly, effects were smaller when the voters were explicitly informed of the selection rule under which poll information was revealed, but the treatment effect was still politically significant and very consistent. Overall, the empirical results from E1 and E2 show limited evidence that the repeated opportunities for learning allowed voters to understand the systematic bias and account for it. The evidence from E3 indicates that it is especially the second part of this statement that matters (failing to account for the bias once one realises it).

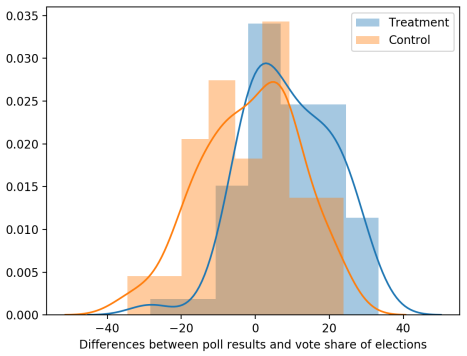
A possible explanation is the genuinely complex environment where voting takes place. For instance, as Figure 4.7a indicates, pure feedback alone is unlikely to be sufficient for learning. Therefore, in E1, in terms of comparing election results to the average revealed poll predictions, the biased treatment condition would not appear as particularly more 'suspicious' to an active learner than the unbiased control condition. It seems that the self-confirming nature of polls renders corrective feedback difficult. Accordingly, voters who are unable to form inferences regarding the strategic nature of the interaction, but only learn from experience, are unable to adjust their behaviour. However, in E3 subjects have a priori information about the bias, and Figure 4.7c indicates that, in the treatment condition, elections tend to diverge from average poll predictions in a systematic way. Despite all this, subjects fail to sufficiently discount the revealed poll information in forming their beliefs (as evidenced by Figure 4.6d). This indicates that even perfect a priori information in conjunction with subsequent feedback are not enough to assist subjects in discounting the results of biased polls appropriately.

How applicable to real-world settings can results derived from our experimental environment be? We believe that our primary result, that people are unable to account for biased polls and hence such bias might robustly distort elections, is likely to generalise to the real world. Our subjects participate in fifteen elections. The number of rounds

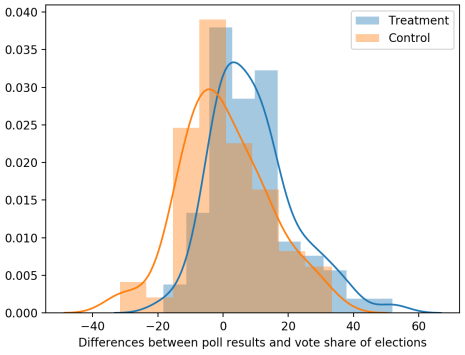
Figure 4.7: Distribution of differences in the vote share of K: revealed poll results vs. elections



(a) E1



(b) E2



(c) E3

Notes. The figures present the distribution of differences between average revealed poll results (presented as K's voting shares) and the actual election results in the same voting round. Data are pooled across rounds and sessions of an experimental condition. Experimental conditions where these differences tend to be large are conditions where subjects have the opportunity to infer that polls are biased.

is reasonably high given the length of the typical experiment in the literature. Indeed, the number of elections that real-life voters may participate in their lifetime is not very large. If anything, the time delay in real election environments might make learning more challenging. Obviously, our stylised environment simplifies important aspects of real elections. However, it seems to us that if subjects are unable to adjust to systematic bias in a stylised environment such as this, they are unlikely to do so in more complicated real elections environments. In addition, the results of E3 indicate that even intergenerational transmission of information about the incentives, agendas and biases of information providers is unlikely to undo the electoral effects of selection and bias in revealed poll results.

In summary, it seems that both our lab and our field evidence raise considerable concerns regarding the risk that biased mechanisms of propagation of poll results affect

democratic outcomes. More evidence from the lab and the field is needed (including replications of this study) before safe policy conclusions can be made. The stakes for electoral policy are particularly high.

Appendix C

C.1 Experimental Instructions

♢ Note that the instructions differ across conditions only in the section **Information about Poll Results**

Instructions

In this study you will be interacting with a fixed group of fourteen other participants for a number of rounds. In each round, the fifteen participants will have the opportunity to vote in an election. The study will consist of 3 practice rounds and 15 regular decision-making rounds. Your performance in the regular rounds counts towards your final earned amount, while practice rounds do not count. For each round, the sequence of actions is illustrated below. In every step, new information will appear at the top of the screen, so please have a look at it carefully before you make any decision or proceed to the next step.

Figure 4.3 here

In each period, you will have the opportunity to vote in an election. One candidate is of PARTY K and the other one is of PARTY J. *Your payoff in each round will depend on the distance of your ideological position from the ideological position of the election winner and on the quality of the election winner.*

Ideological Positions

At the start of each period you will be given a 'position number' between 1 and 15. This number affects how you value the positions of the two candidates. The candidate of Party J is in position 6 and the candidate of Party K is in position 10. These positions remain fixed for both candidates for the entirety of the study, but your position may change every period. In any given round, each of the 15 participants in your group takes a different position. So, every round some participant takes position 1, another participant takes position 2, another participant takes position 3, and so on, up to position 15. The distribution of participants to positions changes every round. Your 'ideological score' from the victory of each candidate is equal to 100 points minus 5 times the difference between your position and the candidate's position.

Candidate Quality

For each of the two candidates an integer number has been randomly drawn for every round. The possible values that this number can take are between 1 and 120

and each number is equally likely to be selected. This 'quality number' reflects the competency of the candidate in handling policy matters. The higher the number is the better the quality of the candidate is. A new quality number was randomly redrawn every period for each candidate. Only some participants in each round will have the opportunity to learn its value.

Informed and Uninformed Participants

In every round some participants are told the quality numbers that have been drawn for the two candidates, e.g. "Candidate J's quality is 100 and candidate K's quality is 24." These are the informed participants. The rest of the participants receive no additional information. Who receives this information is determined by the ideological positions. Participants with positions $\{1, 2, 3, 5, 7, 9, 11\}$ are informed. Participants with positions $\{4, 6, 8, 10, 12, 13, 14, 15\}$ are uninformed. This fact does not change across rounds.

Payoff Example

For example, assume that in a particular round you are in position 3, K's quality in this round is 75 and J's quality in this round is 13. Since candidate K takes position 10, your 'ideological payoff' from K's victory in this round is: $100 - 5 * |3 - 10| = 65$. You also earn an additional score equal to the winner's quality. So, if candidate K wins the election then your total payoff is: $65 + 75 = 140$. On the other hand, Candidate J has position 6. Then, if candidate J wins the election then your total payoff is: $100 - 5 * |3 - 6| + 13 = 85 + 13 = 98$. Please notice that if the difference in the quality between the two candidates in a given round is greater than 20, then you will always receive a higher payoff if the candidate with the higher quality wins, regardless of your ideological position.

Figure 4.2 here

Polls

After the 'informed voters' receive their information, five polling companies will conduct voting intention polls. In each poll, four out of the fifteen participants will be randomly chosen to state their voting preferences. *This means that you may be asked to state your voting intention by one polling company, or by many, or by none.* If you are contacted by many companies, you only have to state your answer once, and the same answer will be used for all of them. Notice that at the time that polls take place, seven voters are informed of the actual quality of the two candidates in the forthcoming election and eight voters are uninformed.

♢ **Information about Poll Results** - *All five polls are revealed (Control condition in E1 and E3)*

After polls have taken place, the findings of the five companies will be revealed. All participants will observe the fraction of votes that each of the two candidates received in the polls of these five companies.

♢ **Information about Poll Results** - *Two biased polls are revealed (Treatment condition in E1 and E2)*

After polls have taken place, the findings of two companies will be revealed. All participants will observe the fraction of votes that each of the two candidates received in the polls of these two companies.

♢ **Information about Poll Results** - *Two out of the five polls are randomly revealed. Subjects are a priori informed about this (Control condition in E2)*

After polls have taken place, the findings of two companies will be revealed to you. These two companies will be selected randomly out of all five that conducted polls. All participants will observe the fraction of votes that each of the two candidates received in the polls of these two companies.

♢ **Information about Poll Results** - *Two biased polls are revealed. Subjects are a priori informed about this (Treatment condition in E3)*

After polls have taken place in each round, the findings of the two companies which exhibit the greatest support for candidate K will be revealed to you. All participants will observe the fraction of votes that each of the two candidates received in the polls of these two companies. For example, consider some illustrative poll results for the five polling companies (A to E), in the following table. If those were the results of all five polls in a given round, then only the results of companies C and E would be revealed to you in that round. If there are ties, these will be broken with a random draw.

COMPANY	A	B	**C**	D	**E**
Candidate K	34%	50%	**100%**	50%	**75%**
Candidate J	66%	50%	**0%**	50%	**25%**

Beliefs about Election Results

After the poll results have been announced to you, and before elections take place, you will be asked to state the vote share that you expect each candidate to receive in the upcoming elections.

Elections and Payoffs

In the end of each period, elections take place, where each participant may vote or abstain. The winner of the election is determined by simple majority. In case of a draw, each candidate receives an equal chance of being selected as the winner.

Round Payoffs and Aggregate Payoffs from the Study

As described earlier, your total payoff from each round is the sum of the winner's quality and your 'ideological payoffs' from the candidate's victory. *Your total payoffs from this study will be the sum of all payoffs that you accumulate in each of the 15 regular rounds, plus your participation fee.* They will be paid to you in cash, at the end of the study. Each earned point will correspond to **half a penny**. You will now participate in three practice rounds. If you have any questions, please raise your hand and your question will be addressed individually.

C.2 Additional Graphs

Figure C.1: Poll outcomes in treatment sessions of E1 (T1-T4)

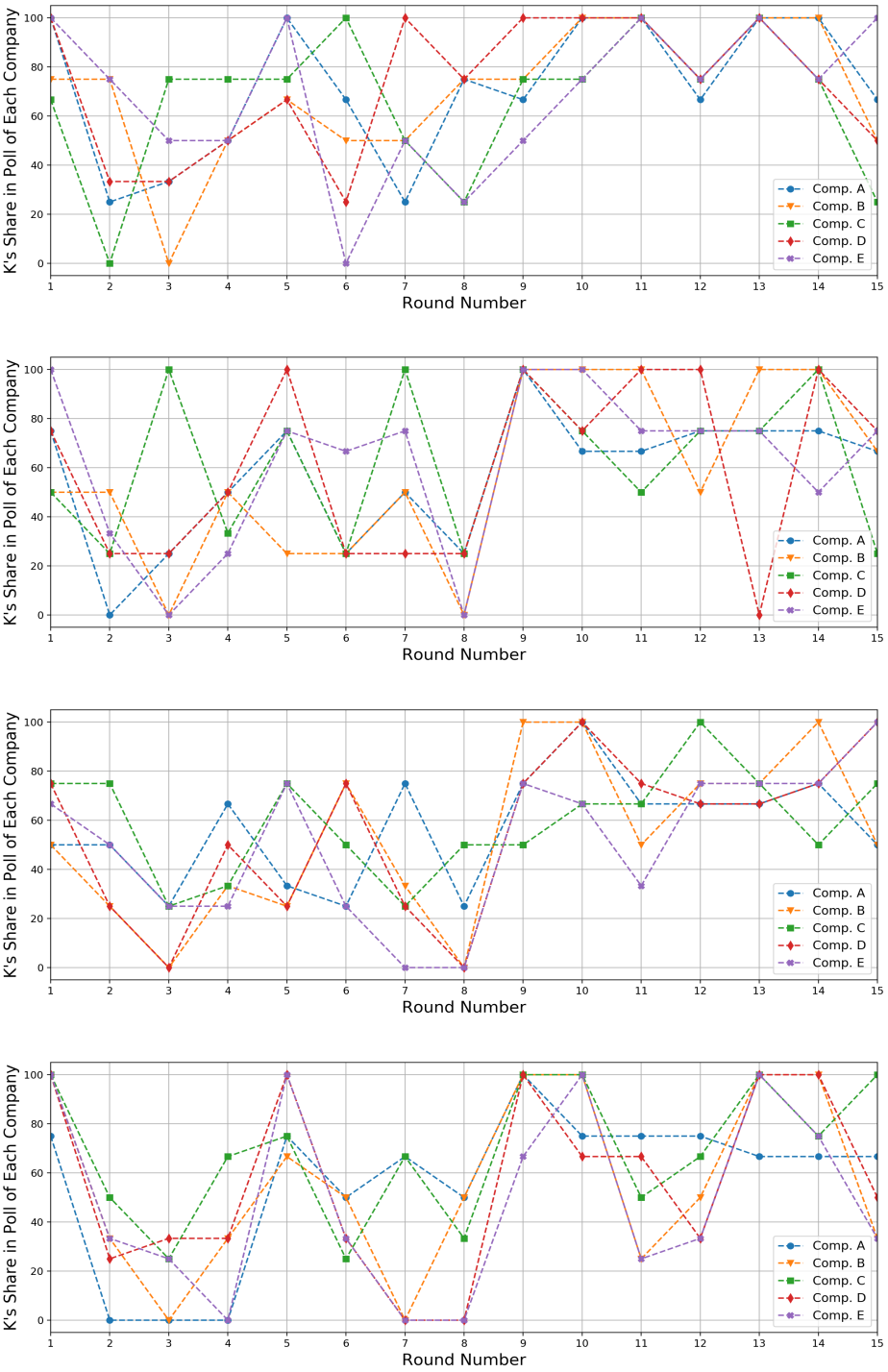


Figure C.2: Average beliefs vs. poll outcomes in control sessions of E1 (C1-C4)

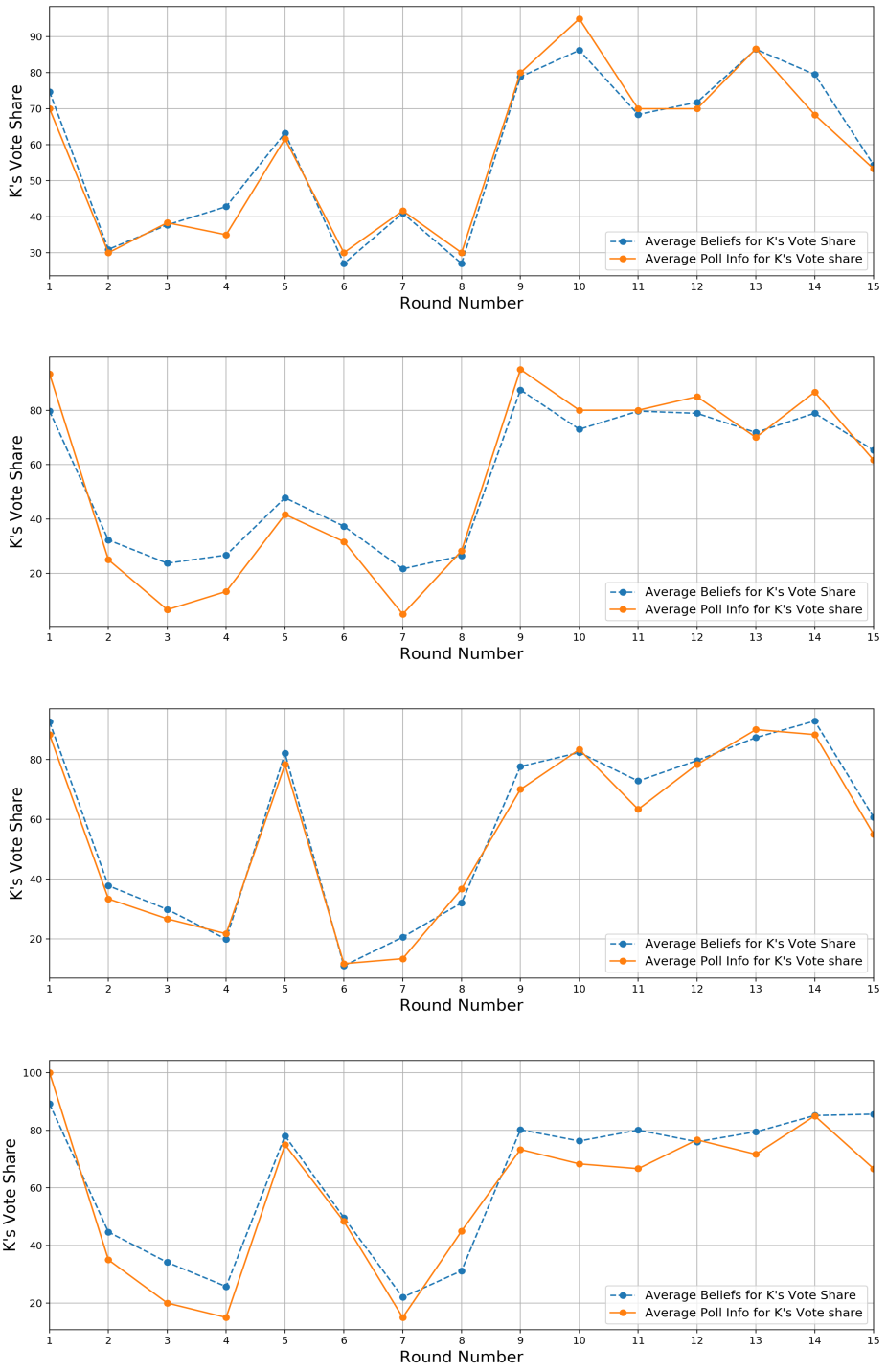


Figure C.3: Average beliefs vs. poll outcomes in treatment sessions of E1 (T1-T4)

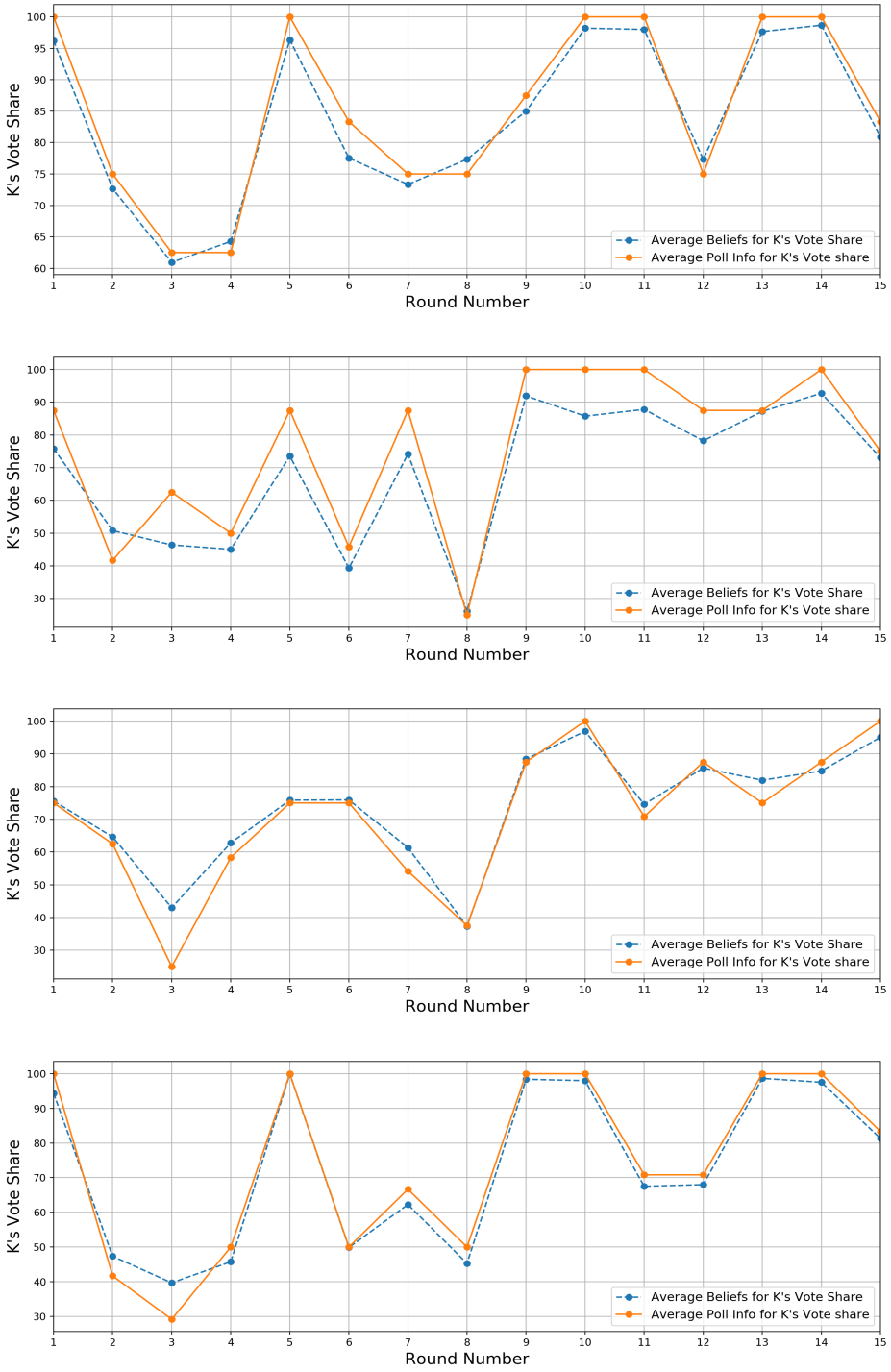


Figure C.4: Poll outcomes in treatment sessions of E2 (T1-T4)

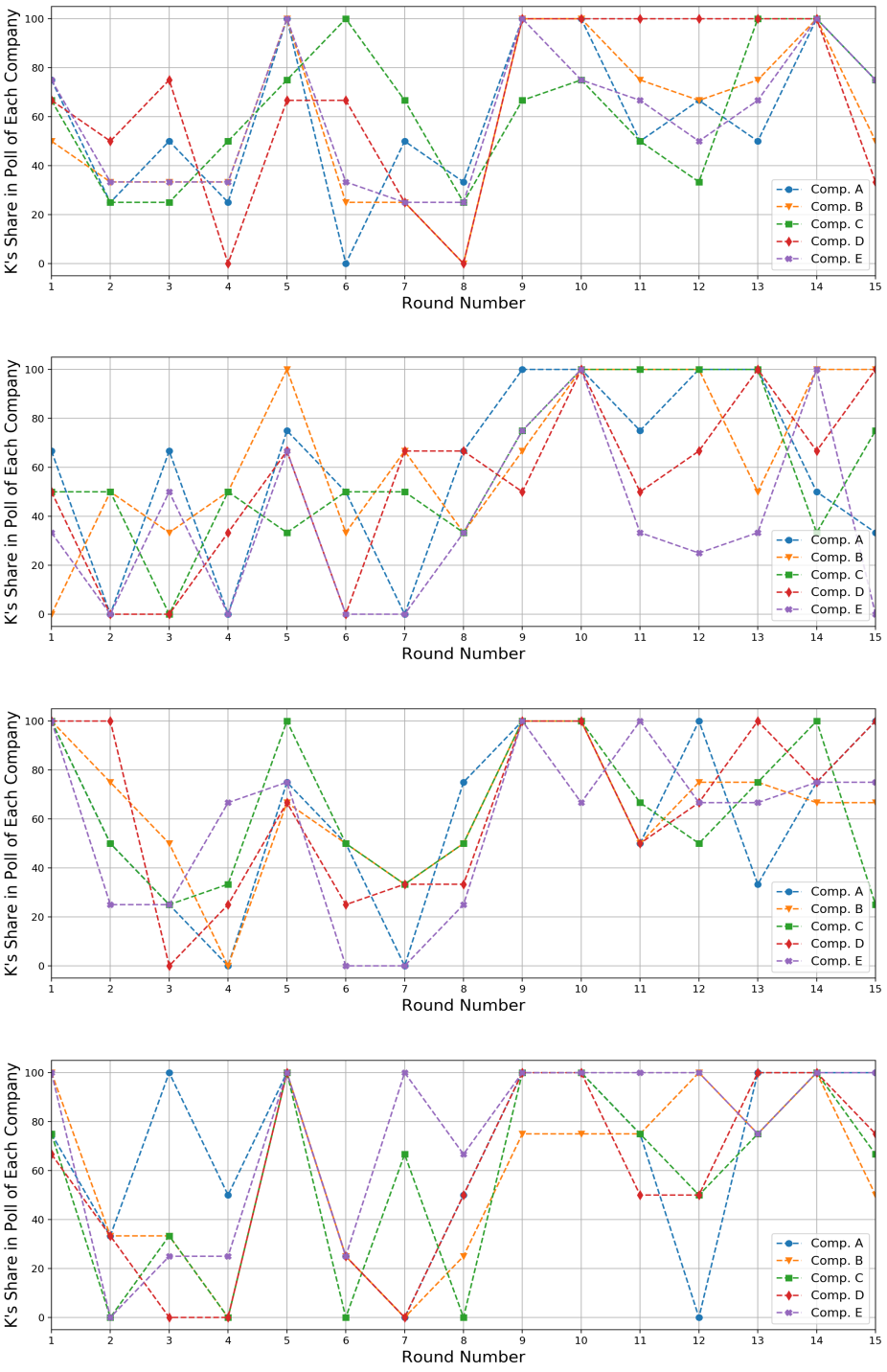


Figure C.5: Average beliefs vs. poll outcomes in control sessions of E2 (C1-C4)

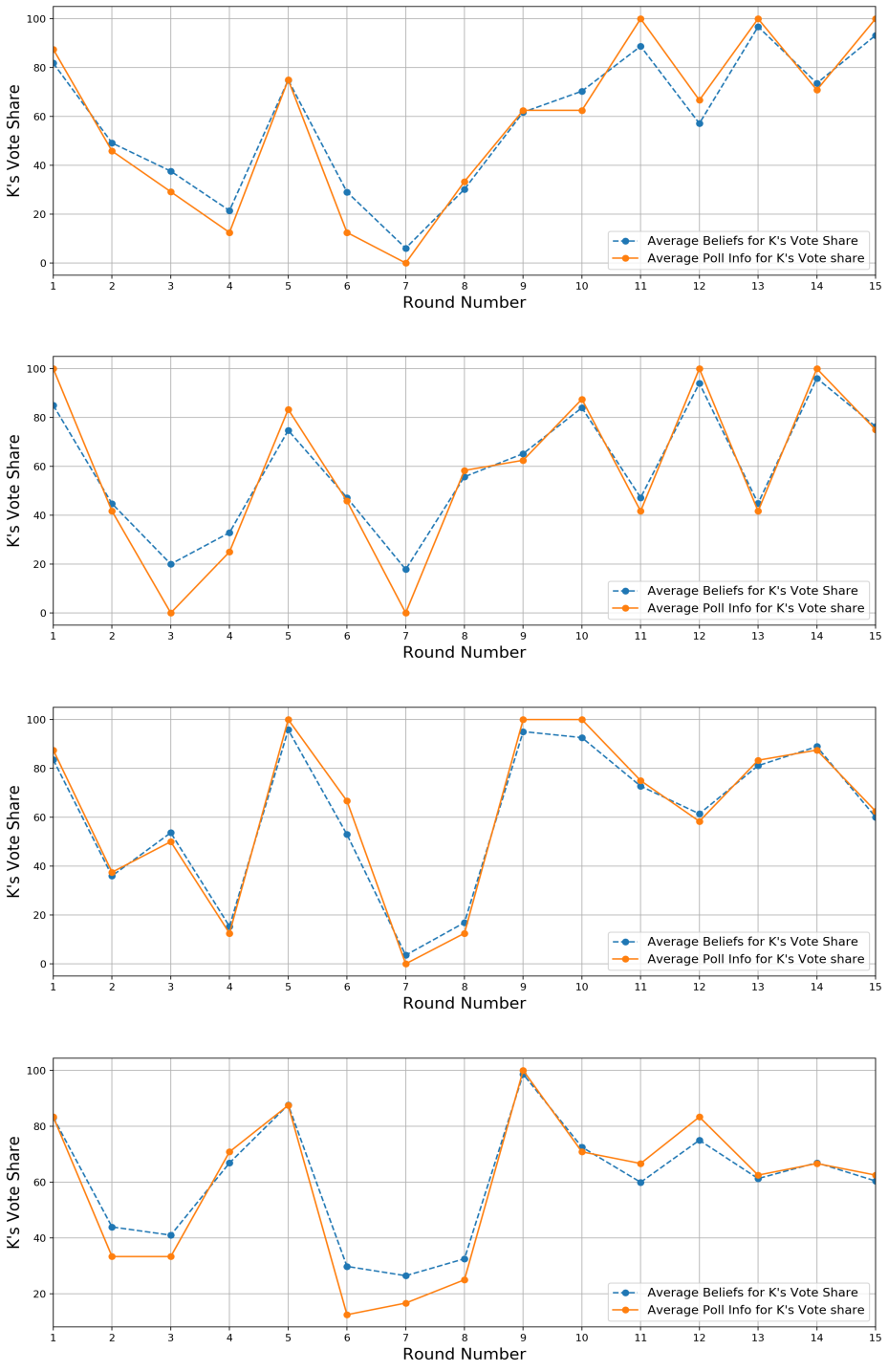


Figure C.6: Average beliefs vs. poll outcomes in treatment sessions of E2 (T1-T4)

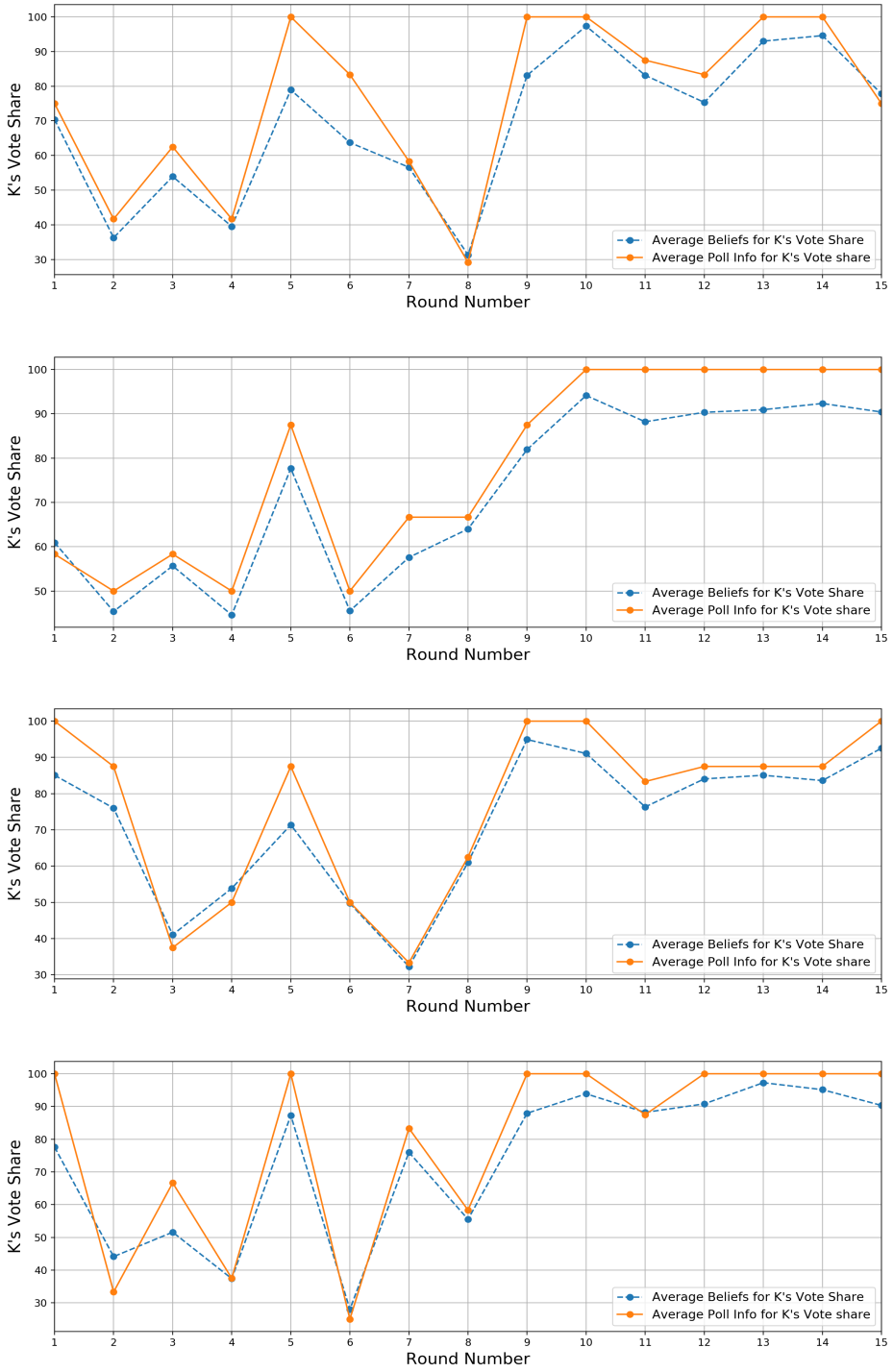


Figure C.7: Poll outcomes in treatment sessions of E3 (T1-T5)

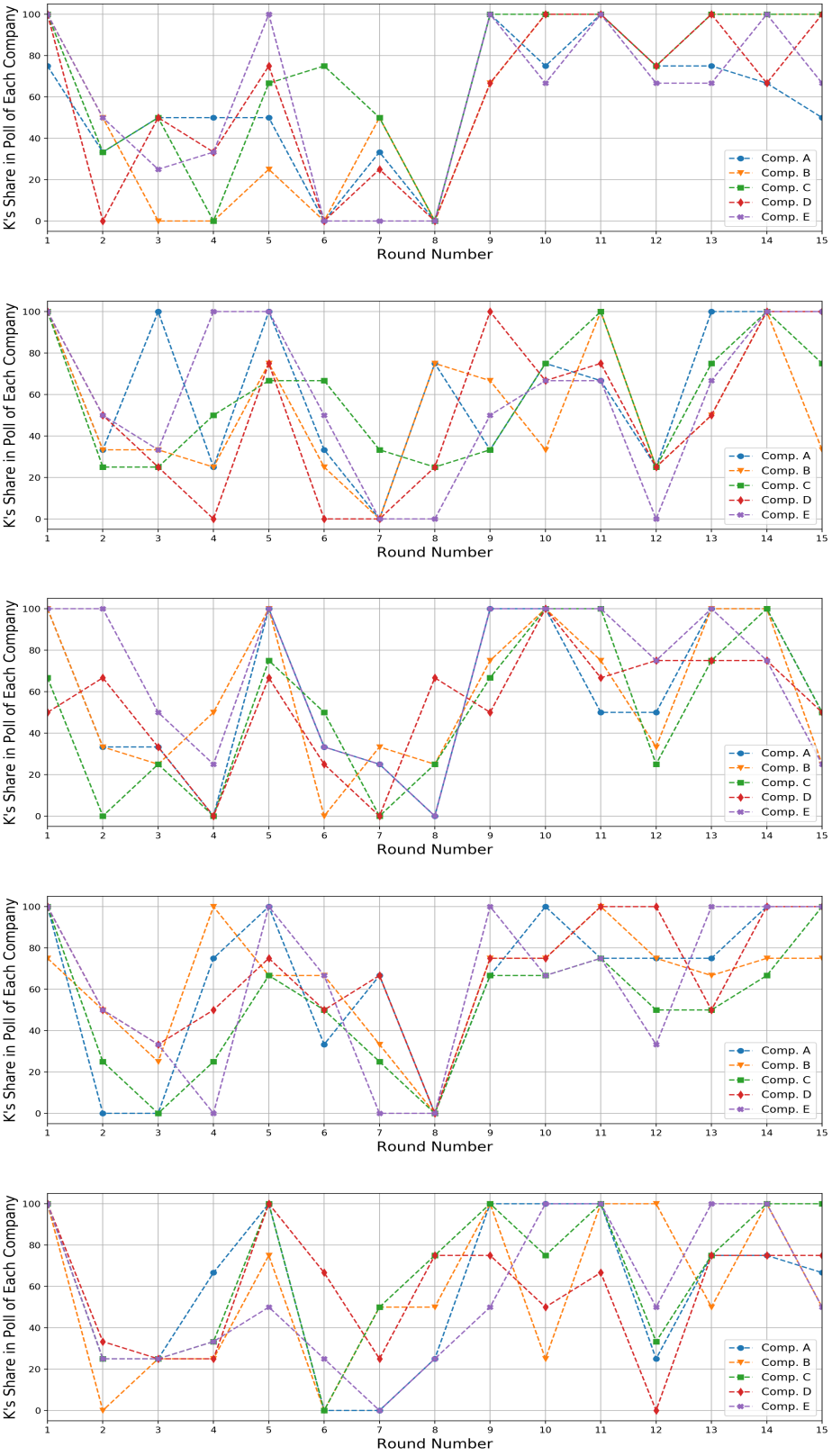


Figure C.8: Average beliefs vs. poll outcomes in control sessions of E3 (C1-C4)

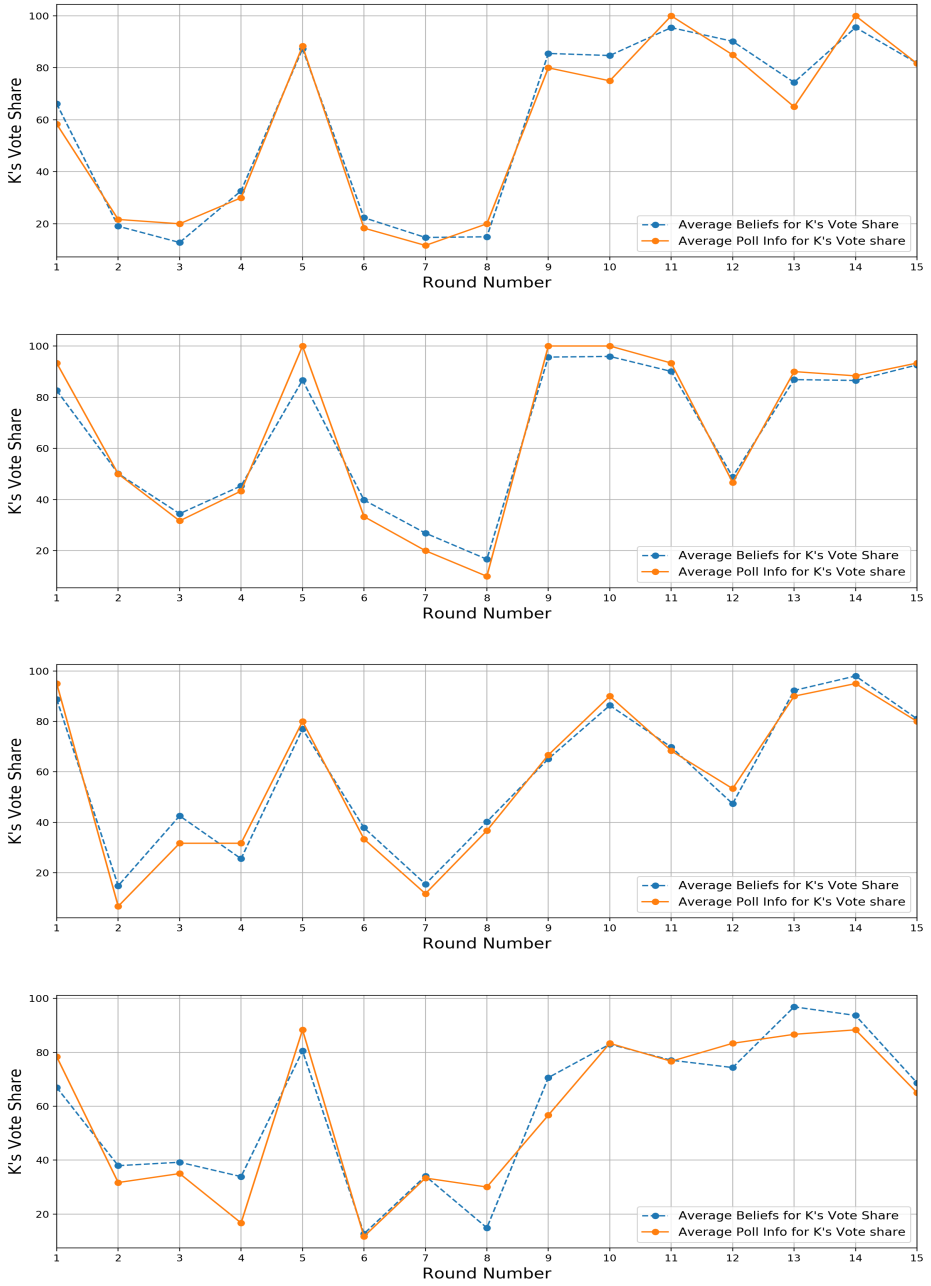
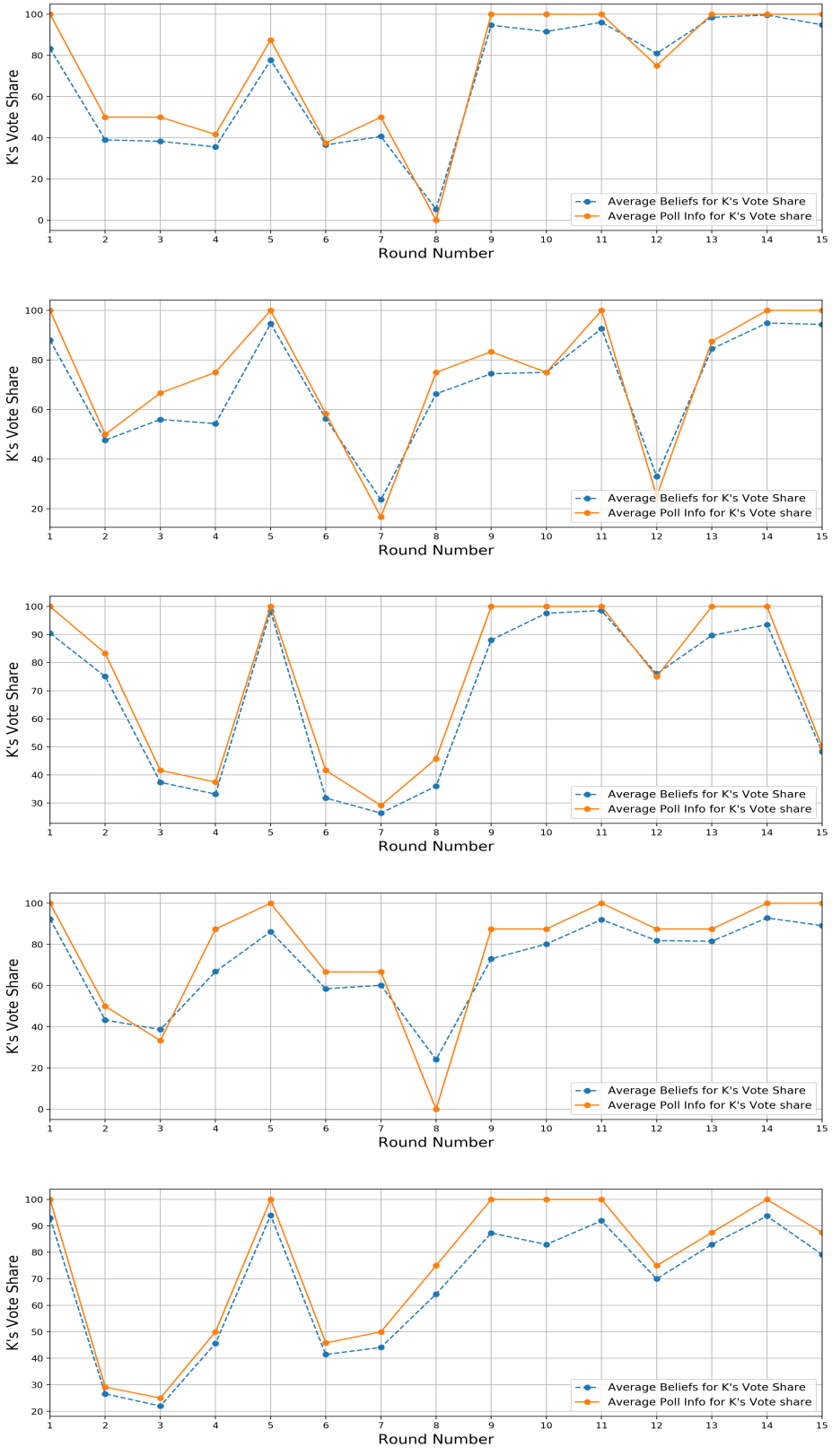


Figure C.9: Average beliefs vs. poll outcomes in treatment sessions of E3 (T1-T5)



C.3 Additional Tables

Table C.1: Descriptive summary of voting behaviour at the poll stage, pooled at session level, E1

	session	E1_C1	E1_C2	E1_C3	E1_C4	E1_T1	E1_T2	E1_T3	E1_T4
uninformed	J	26.09%	29.21%	29.67%	38.30%	17.20%	30.85%	46.67%	21.74%
	K	42.39%	47.19%	37.36%	51.06%	67.74%	53.19%	43.33%	47.83%
	N	31.52%	23.60%	32.97%	10.64%	15.05%	15.96%	10.00%	30.43%
informed	J	44.05%	42.50%	44.44%	43.82%	34.94%	45.24%	38.37%	51.19%
	K	55.95%	50.00%	54.32%	55.06%	62.65%	54.76%	59.30%	47.62%
	N	0.00%	7.50%	1.23%	1.12%	2.41%	0.00%	2.33%	1.19%

Table C.2: Descriptive summary of voting behaviour at the poll stage, pooled at session level, E2

	session	E2_C1	E2_C2	E2_C3	E2_C4	E2_T1	E2_T2	E2_T3	E2_T4
uninformed	J	30.21%	29.21%	33.71%	22.92%	28.71%	26.32%	27.37%	29.21%
	K	43.75%	56.18%	39.33%	51.04%	48.51%	37.89%	47.37%	41.57%
	N	26.04%	14.61%	26.97%	26.04%	22.77%	35.79%	25.26%	29.21%
informed	J	38.75%	45.45%	44.19%	46.91%	45.56%	46.84%	37.50%	34.94%
	K	57.50%	53.41%	55.81%	50.62%	48.89%	50.63%	62.50%	61.45%
	N	3.75%	1.14%	0.00%	2.47%	5.56%	2.53%	0.00%	3.61%

Table C.3: Descriptive summary of voting behaviour at the poll stage, pooled at session level, E3

	session	E3_C1	E3_C2	E3_C3	E3_C4	E3_T1	E3_T2	E3_T3	E3_T4	E3_T5
uninformed	J	34.88%	21.98%	29.03%	23.76%	30.85%	20.83%	27.84%	23.96%	32.97%
	K	50.00%	43.96%	41.94%	39.60%	38.30%	43.75%	47.42%	54.17%	56.04%
	N	15.12%	34.07%	29.03%	36.63%	30.85%	35.42%	24.74%	21.88%	10.99%
informed	J	43.82%	37.18%	43.02%	48.10%	37.04%	52.33%	44.83%	42.35%	44.57%
	K	49.44%	60.26%	56.98%	50.63%	62.96%	47.67%	54.02%	56.47%	54.35%
	N	6.74%	2.56%	0.00%	1.27%	0.00%	0.00%	0.00%	1.18%	1.09%

Notes. 'N' stands for non-participation.

Bibliography

Adaval, R. & Wyer, R. S. (2011). Conscious and nonconscious comparisons with price anchors: Effects on willingness to pay for related and unrelated products. *Journal of Marketing Research*, *48*(2), 355–365.

Agranov, M., Goeree, J. K., Romero, J., & Yariv, L. (2017). What makes voters turn out: The effects of polls and beliefs. *Journal of the European Economic Association*, *16*(3), 825–856.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Alevy, J. E., Landry, C. E., & List, J. A. (2015). Field experiments on the anchoring of economic valuations. *Economic Inquiry*, *53*(3), 1522–1538.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, *21*(4), 503–546.

Allcott, H. & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–36.

Andrersson, P. & Wisaeus, B. (2013). Age and anchoring. http://arc.hhs.se/download.aspx?MediumId=1927.

Araña, J. E. & León, C. J. (2008). Do emotions matter? coherent preferences under anchoring and emotional effects. *Ecological Economics*, *66*(4), 700–711.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106.

Bavolár, J. (2017). Experience with the product does not affect the anchoring effect, but the relevance of the anchor increases it. *Ekonomickỳ časopis*, *95*(03), 282–293.

Bergman, O., Ellingsen, T., Johannesson, M., & Svensson, C. (2010). Anchoring and cognitive ability. *Economics Letters*, *107*(1), 66–68.

Blankenship, K. L., Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *Journal of Experimental Social Psychology*, *44*(6), 1465–1476.

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, *71*, 117–120.

Boeri, T., Mishra, P., Papageorgiou, C., & Spilimbergo, A. (2018). A dialogue between a populist and an economist. In *AEA Papers and Proceedings*, volume 108, (pp. 191–95).

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, *127*(3), 1243–1285.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013a). Salience and asset prices. *The American Economic Review*, *103*(3), 623–628.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013b). Salience and consumer choice. *Journal of Political Economy*, *121*(5), 803–843.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2015). Salience theory of judicial decisions. *The Journal of Legal Studies*, *44*(S1), S7–S33.

Borenstein, M., Cooper, H., Hedges, L., & Valentine, J. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, *2*, 221–235.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Is the internet causing political polarization? evidence from demographics. Technical report, NBER Working Paper No. w23258.

Brown, K. M. & Zech, C. E. (1973). Welfare effects of announcing election forecasts. *Public Choice*, *14*(1), 117–123.

Brzozowicz, M. & Krawczyk, M. (2019). Anchors don't hold for real? the anchoring effect and hypothetical bias in declared willingness to pay. *Unpublished, in review.*

Brzozowicz, M., Krawczyk, M., Kusztelak, P., et al. (2017). Do anchors hold for real? anchoring effect and hypothetical bias in declared wtp. *University of Warsaw, Faculty of Economic Sciences.*

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference.* Springer.

Camerer, C. F. & Loewenstein, G. (2003). *Behavioral economics: Past, present, future.* Princeton University Press.

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, *99*(4), 1145–77.

Cipullo, D. & Reslow, A. (2019). *Biased Forecasts to Affect Voting Decisions?: The Brexit Case*. Uppsala University.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

DellaVigna, S. & Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, *122*(3), 1187–1234.

Dertwinkel-Kalt, M. & Köster, M. (2019). Salience and skewness preferences. *Journal of the European Economic Association*. https://doi.org/10.1093/jeea/jvz035.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 11–33.

Dogerlioglu-Demir, K. & Koçaş, C. (2015). Seemingly incidental anchoring: the effect of incidental environmental anchors on consumers' willingness to pay. *Marketing Letters*, *26*(4), 607–618.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634.

Epley, N. & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199–212.

Feddersen, T. & Pesendorfer, W. (1997). Voting behavior and information aggregation in elections with private information. *Econometrica*, *65*, 1029–1058.

Fleitas, D. W. (1971). Bandwagon and underdog effects in minimal-information elections. *American Political Science Review*, *65*(2), 434–438.

Forsythe, R., Rietz, T., Myerson, R., & Weber, R. (1996). An experimental study of voting rules and polls in three-candidate elections. *International Journal of Game Theory*, *25*(3), 355–383.

Frydman, C. & Mormann, M. M. (2018). The role of salience in choice under risk: An experimental investigation. *Available at SSRN 2778822*.

Fudenberg, D., Levine, D. K., & Maniadis, Z. (2012). On the robustness of anchoring effects in wtp and wta experiments. *American Economic Journal: Microeconomics*, *4*(2), 131–145.

Funke, M., Schularick, M., & Trebesch, C. (2016). Going to extremes: Politics after financial crises, 1870–2014. *European Economic Review*, *88*, 227–260.

Gentzkow, M. & Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, *78*(1), 35–71.

Georgalos, K. (2019). An experimental test of the predictive power of dynamic ambiguity models. *Journal of Risk and Uncertainty*. forthcoming.

Gerber, A., Hoffman, M., Morgan, J., & Raymond, C. (2017). One in a million: Field experiments on perceived closeness of the election and voter turnout. Technical report, NBER Working Paper No. w23071.

Goeree, J. K. & Grosser, J. (2007). Welfare reducing polls. *Economic Theory*, *31*(1), 51–68.

Gonzalez, R. & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Green, D., Jacowitz, K. E., Kahneman, D., & McFadden, D. (1998). Referendum contingent valuation, anchoring, and willingness to pay for public goods. *Resource and Energy Economics*, *20*(2), 85–116.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, *1*(1), 114–125.

Grether, D. M. & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review*, *69*(4), 623–638.

Guriev, S. (2018). Economic drivers of populism. In *AEA Papers and Proceedings*, volume 108, (pp. 200–203).

Hargreaves-Heap, S. & Clark, C. G. (2017). *Economic Man*, (pp. 1–4). London: Palgrave Macmillan UK.

Hey, J. D., Lotito, G., & Maffioletti, A. (2010). The descriptive and predictive adequacy of theories of decision making under uncertainty/ambiguity. *Journal of Risk and Uncertainty*, *41*(2), 81–111.

Hey, J. D. & Pace, N. (2014). The explanatory and predictive power of non two-stage-probability theories of decision making under ambiguity. *Journal of Risk and Uncertainty*, *49*(1), 1–29.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2019). Cochrane handbook for systematic reviews of interventions. `https://training.cochrane.org/handbook/current`.

Higgins, J. P. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539–1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj*, *327*(7414), 557–560.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 297–307.

Inglehart, R. & Norris, P. (2017). Trump and the populist authoritarian parties: the silent revolution in reverse. *Perspectives on Politics*, *15*(2), 443–454.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), 696–701.

Iyengar, S. (1991). *Is Anyone Responsible?: How Television Frames Political Issues*. The University of Chicago Press.

Jacowitz, K. E. & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166.

Jung, M. H., Perfecto, H., & Nelson, L. D. (2016). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. *Journal of Marketing Research*, *53*(3), 354–368.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, *5*(1), 193–206.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Kahneman, D. & Tversky, A. (1981). The simulation heuristic. Technical report, Stanford University.

Kaltwasser, C. (2018). Studying the (economic) consequences of populism. In *AEA Papers and Proceedings*, volume 108, (pp. 204–07).

Klor, E. F. & Winter, E. (2007). The welfare effects of public opinion polls. *International Journal of Game Theory*, *35*(3), 379.

Koçaş, C. & Dogerlioglu-Demir, K. (2014). An empirical investigation of consumers' willingness-to-pay and the demand function: The cumulative effect of individual differences in anchored willingness-to-pay responses. *Marketing Letters*, *25*(2), 139–152.

Königsheim, C., Lukas, M., & Nöth, M. (2019). Salience theory: Calibration and heterogeneity in probability distortion. *Journal of Economic Behavior & Organization*, *157*, 477–495.

Kontek, K. (2016). A critical note on salience theory of choice under risk. *Economics Letters*, *149*, 168–171.

Kothiyal, A., Spinu, V., & Wakker, P. P. (2014). An experimental test of prospect theory for predicting choice under ambiguity. *Journal of Risk and Uncertainty*, *48*(1), 1–17.

Larsen, E. G. & Fazekas, Z. (2019). Transforming stability into change: How the media select and report opinion polls. In *The International Journal of Press/Politics*.

LeBoeuf, R. A. & Shafir, E. (2009). Anchoring on the" here" and" now" in time and distance judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 81.

Li, T., Fooks, J. R., Messer, K. D., & Ferraro, P. J. (2019). A field experiment to estimate the effects of anchoring and framing on residents' willingness to purchase water runoff management technologies. *Resource and Energy Economics*. In press.

Lichtenstein, S. & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(1), 46.

Maniadis, Z. (2014). Selective revelation of public information and self-confirming equilibrium. *International Journal of Game Theory*, *43*(4), 991–1008.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, *104*(1), 277–290.

Matthews, J. S., Pickup, M., & Cutler, F. (2012). The mediated horserace: Campaign polls and poll reporting. *Canadian Journal of Political Science*, *45*(2), 261–287.

McKelvey, R. D. & Ordeshook, P. C. (1984). Rational expectations in elections: Some experimental results based on a multidimensional model. *Public Choice*, *44*(1), 61–102.

McKelvey, R. D. & Ordeshook, P. C. (1985). Elections with limited information: A fulfilled expectations model using contemporaneous poll and endorsement data as information sources. *Journal of Economic Theory*, *36*(1), 55–85.

Meffert, M. F. & Gschwend, T. (2011). Polls, coalition signals and strategic voting: An experimental investigation of perceptions and effects. *European Journal of Political Research*, *50*(5), 636–667.

Mellon, J. & Prosser, C. (2017). Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, *4*(3), 1–9.

Morwitz, V. G. & Pluzinski, C. (1996). Do polls reflect opinions or do opinions reflect polls? the impact of political polling on voters' expectations, preferences, and behavior. *Journal of Consumer Research*, *23*(1), 53–67.

Mussweiler, T. & Strack, F. (2001). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition*, *19*(2), 145–160.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142–1150.

Nielsen, C. S., Sebald, A. C., & Sørensen, P. N. (2018). Testing for salience effects in choices under risk. Technical report, Working Paper, University of Copenhagen.

Northcraft, G. B. & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and human decision processes*, *39*(1), 84–97.

Nunes, J. C. & Boatwright, P. (2004). Incidental prices and their effect on willingness to pay. *Journal of Marketing Research*, *41*(4), 457–466.

Oesch, D. (2008). Explaining workers' support for right-wing populist parties in western europe: Evidence from austria, belgium, france, norway, and switzerland. *International Political Science Review*, *29*(3), 349–373.

Palfrey, T. R. (2016). Experiments in political economy. *Handbook of Experimental Economics*, *2*.

Peterson, R. A. & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175.

Petrova, M. (2008). Inequality and media capture. *Journal of Public Economics*, *92*(1-2), 183–212.

Plott, C. R. (1982). A comparative analysis of direct democracy, two candidate elections, and three candidate elections in an experimental environment. Technical report, California Institute of Technology Working Paper Series.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Rodrik, D. (2018). Is populism necessarily bad economics? In *AEA Papers and Proceedings*, volume 108, (pp. 196–199).

Rothschild, D. & Malhotra, N. (2014). Are public opinion polls self-fulfilling prophecies? *Research & Politics*, *1*(2), 1–10.

Schläpfer, F. & Schmitt, M. (2007). Anchors, endorsements, and preferences: a field experiment. *Resource and Energy Economics*, *29*(3), 229–243.

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*(522), 607–614.

Simon, H. A. (1957). *Models of man; social and rational.* Wiley.

Simonson, I. & Drolet, A. (2004). Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of consumer research*, *31*(3), 681–690.

Sinclair, B. & Plott, C. R. (2012). From uninformed to informed choices: Voters, pre-election polls and updating. *Electoral Studies*, *31*(1), 83–95.

Song, F., Easterwood, A., Guilbody, S., Duley, L., & Sutton, A. (2000). Publication and other selection bias in systematic reviews. *Health Technology Assessment*, *4*(10), 1–115.

Soroka, S. N. (2014). *Negativity in Democratic Politics: Causes and Consequences.* Cambridge University Press.

Strack, F. & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, *73*(3), 437.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B., & Smith, P. (2016). Report of the inquiry into the 2015 british general election opinion polls. In *NCRM, British Polling Council, Market Research Society*.

Sugden, R., Zheng, J., & Zizzo, D. J. (2013). Not all anchors are created equal. *Journal of Economic Psychology*, *39*, 21–31.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods*, *7*(1), 13–26.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Tanford, S., Choi, C., & Joe, S. J. (2019). The influence of pricing strategies on willingness to pay for accommodations: Anchoring, framing, and metric compatibility. *Journal of Travel Research*, *58*(6), 932–944.

Thompson, S. G. & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine*, *21*(11), 1559–1573.

Tufano, F. (2010). Are 'true'preferences revealed in repeated markets? an experimental demonstration of context-dependent valuations. *Experimental Economics*, *13*(1), 1–13.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Van Hauwaert, S. M. & Van Kessel, S. (2018). Beyond protest and discontent: A cross-national analysis of the effect of populist attitudes and issue positions on populist party support. *European Journal of Political Research*, *57*(1), 68–92.

Veblen, T. (1898). Why is economics not an evolutionary science?

Von Neumann, J. & Morgenstern, O. (1947). *Theory of games and economic behavior, 2nd rev.* Princeton University Press.

Wansink, B., Kent, R. J., & Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research*, *35*(1), 71–81.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, *37*(1), 62–69.

Whiteley, P. (2016). Why did the polls get it wrong in the 2015 general election? evaluating the inquiry into pre-election polls. *The Political Quarterly*, *87*(3), 437–442.

Whyte, G. & Sebenius, J. K. (1997). The effect of multiple anchors on anchoring in individual and group judgment. *Organizational behavior and human decision processes*, *69*(1), 75–85.

Wilson, D. B. (2001). Practical meta-analysis effect size calculator [online calculator]. https://campbellcollaboration.org/research-resources/effect-size-calculator.html.

Wlezien, C., Jennings, W., Fisher, S., Ford, R., & Pickup, M. (2013). Polls and the vote in britain. *Political Studies*, *61*, 66–91.

Wu, C., Cheng, F., & Lin, H. (2008). Exploring anchoring effect and the moderating role of repeated anchor in electronic commerce. *Behaviour & Information Technology*, *27*(1), 31–42.

Wu, C.-S. & Cheng, F.-F. (2011). The joint effect of framing and anchoring on internet buyers' decision-making. *Electronic Commerce Research and Applications*, *10*(3), 358–368.

Wu, G. & Markle, A. B. (2008). An empirical test of gain-loss separability in prospect theory. *Management Science*, *54*(7), 1322–1335.

Yoon, S. & Fong, N. (2019). Uninformative anchors have persistent effects on valuation judgments. *Journal of Consumer Psychology*, *29*, 391–410.

Yoon, S., Fong, N. M., & Dimoka, A. (2013). The robustness of anchoring effects on market good valuations. *Available at SSRN 2352692*.

Yu, L., Gao, Z., Sims, C., & Guan, Z. (2017). Effect of price on consumers' willingness to pay: is it from quality perception or price anchoring? *Agricultural & Applied Economics Association Annual Meeting, Chicago, Illinois*.