Essays in Behavioural Economics: Representation and Weighting of Numerical Information

A thesis submitted to the University of Dublin, Trinity College

In application for the degree of Doctor of Philosophy

by

Féidhlim McGowan

Supervised by: Prof. Eleanor Denny and Prof. Pete Lunn

Head of School: Prof. Carol Newman

2022

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Féidhlim McGowan

Summary

This thesis consists of four essays. The first two essays (Chapters 2 and 3) present experimental tests of how numerical information is represented in the brain, and the implications of this representation for economic decisions. The latter two chapters use field evidence to infer how decision makers weight numerical information in economic choices that include a social dimension.

Chapter 2 experimentally investigates systematic bias in intuitive summation. The traditional (implicit) assumption in economics is that rational agents can add a sequence of numbers without error. Or, if errors occur, that they are random and hence cancel out in aggregate. Instead of random error, we find consistent underestimation, even though accuracy is incentivised, and the task is framed to be familiar and economically meaningful. Underestimation occurs both when people are asked to generate a best guess for the sum, and when asked to compare their impression of the sum to a given number. Underestimation bias can help explain anomalies in consumer choice.

Chapter 3 tests the explanatory power of underestimation bias, and a limited-attention phenomenon called concentration bias, on willingness to pay for energy-efficient investments. These cognitive mechanisms can underlie behaviour generally attributed to discounting. In a pre-registered experiment on a large, nationally representative sample of car buyers, we elicit willingness-to-pay (WTP) for an improvement in fuel economy. The results support the pre-registered hypotheses and suggest that the proportion of the energy-efficiency gap attributed to time preferences may be exaggerated.

Chapter 4 explores the nature of status-signalling in the market for new cars. Traditional models of signaling assume that all agents are fully attentive. Empirical evidence of limited attention suggests status signals need to be salient and have an obvious meaning to be inferred correctly. In this paper, variation in salience comes

from random differences in the presentation format of age identifiers on licence plates in Ireland and Great Britain. The age identifier is more salient in Ireland. Results indicate that the difference in salience has a causal effect on market demand both in terms of when purchases occur, and the type of cars purchased. Premium makes such as BMW - conventionally status signals - are less popular when the age identifier is more salient, implying substitution between status attributes of car make and age. These findings can inform labelling policy to nudge consumption into a pattern which generates positive externalities.

Chapter 5 investigates how groups of experts weight a quantitative attribute when deciding how to allocate a scarce resource, namely literary prizes. I analyse the population of shortlisted novels for three literary prizes covering a time span of 1963-2021. I show that longer novels are more likely to win. The result is robust to controlling for author gender and Goodreads rating, and to whether one uses absolute page length or relative length on the shortlist. The size of the effect suggests other valid cues are underweighted in the process of selecting a winner. Judgment and decision making research suggests several causes of the bias. One is the representativeness heuristic: Other explanations include an effort heuristic and the effects of accountability in decisions. Heightened sensitivity to magnitude under joint-evaluation is another potential cause. These results may explain previous findings that Booker Prize winners are not higher quality than shortlisted novels. The findings have broader implications for the inferred quality of expert judgment.

Acknowledgements

My deepest thanks first and foremost go to Pete Lunn, my external supervisor. Thank you for being the most brilliant mentor, and a true friend.

I owe a huge debt of gratitude to my internal supervisor, Eleanor Denny, for expert guidance and invaluable support at every step of the PhD. I am especially grateful for your encouragement and reassurance at times when I was doubting myself.

I thank my funder, the Irish Research Council, for their financial support since my PhD began. I am thankful also to TRiSS for the award of a Postgraduate Research Fellowship grant in 2020. I am deeply appreciative to Trinity for the Foundation Scholarship I was awarded in 2013. The material benefits made a postgraduate life more feasible, but, more importantly, it allowed me to make friends for life: Colm, Conor and George, I am looking forward to the reunion already. And part of the reason I came back to Dublin for my PhD was because of the great friends I still have here. To Iarfhlaith, James, and William, I'm glad you're in my life. And to all the others - you know who you are - thanks for your friendship, whether it was over coffee, at a pub quiz, or on a run.

My research journey began as a intern in the ESRI in 2014. I am indebted to fellow Sligo man John O'Hagan for the initial encouragement to go down the research track, and for spurring me on at several points since then. I am grateful to my former colleagues in the wider ESRI for the kind and congenial atmosphere which made a career in research seem like a decent gig.

I am also thankful to the ESRI for allowing me to be a Visiting PhD Student. The access to the Institute pre-pandemic meant I was surrounded by an excellent team of behavioural scientists from whom I could learn. To Cameron, Ciarán, Deirdre, Hannah, Martina, and Shane, thank you for the advice and feedback, and for piloting my experiments.

My student experience has been enriched immensely by being part of DUHAC. Running with friends on College Park has been a source of happiness and belonging for me. It is a remarkable club. In particular, I can't express how grateful I am to our coach, Dr. Iain Morrison, for his empathy and energy, which is a marvel and something to aspire to.

To Conor, I owe you a tremendous debt of gratitude. My only regret looking back over the PhD is that we didn't get more time in House 40. Thanks for all the late-night kitchen chats, all the long runs, and for always being there for me. A million others would kill for a friend like you.

To Lisa, I am eternally grateful for all your love and support. I couldn't have done this without you.

Finally, I cannot express how fortunate I am to have the most wonderful family backing me all the way. Domhnall and Clíodhna, it is very easy to be the middle child when I have siblings like you. To Mum and Dad, thank you for all you've done for me. Mum, you continue to be my inspiration. You encouraged my curiosity from the very start, even when all I was interested in was digging holes in the garden. I dedicate this thesis to you.

Contents

1 Introduction **1**

2 Consistent Underestimation in the Intuitive Summation of Monetary Amounts **9**

2.1 Introduction 10

2.1.1 Numeric Cognition 13

2.1.2 Household Finance and Bill Perceptions 15

2.2 Experiment 1 16

2.2.1 Method 16

2.2.2 Experiment 1 Results 20

2.2.3 Experiment 1 Discussion 28

2.3 Experiment 2 29

2.3.1 Method 30

2.3.2 Experiment 2 Results 32

2.3.3 Forced-Choice Results 35

2.3.4 Estimation Strategies 40

2.3.5 Simulating Choice Data from Judgment Responses 41

2.3.6 Experiment 2 Discussion 43

2.4 General Discussion 44

2.5 Supplementary Material 49

2.5.1 Power Analysis 49

2.5.2 Number Sequence Construction 51

2.5.3 Departures from Preregistration Plan 52

2.5.4 Experiment 1: Additional Analyses 55

2.5.5 Experiment 2: Additional Analyses 62
2.5.6 Experimental Instructions 67

3 Looking beyond time preference: testing cognitive mechanisms relevant to the energy paradox 71

3.1 Introduction 72
3.2 Related Literature and Decision Framework 76
3.2.1 Leading Causes of the Energy-Efficiency Gap 76
3.2.2 Experimental Evidence on Labelling and Energy Valuation . . 78
3.2.3 Concentration Bias and Underestimation Bias 80
3.2.4 Decision Framework 81
3.3 The Experiment 83
3.3.1 Design and Hypotheses 84
3.3.2 Method 88
3.4 Results 89
3.4.1 Fuel Cost-Payment Schedule Results 90
3.4.2 Regression Results 91
3.5 Discussion and Conclusion 98
3.6 Supplementary Material 103

4 Salience, Status and Social Norms: A Natural Experiment on Car Licence Plate Formats 112

4.1 Introduction 113
4.2 Related Literature 117
4.2.1 Status Utility: Theory and Empirics 117
4.2.2 Salience 119
4.2.3 Age Identifiers within Salience Framework 122
4.3 Institutional Context and Data 124
4.3.1 Data 127
4.4 Results 128
4.4.1 Effect of Switching to Bi-annual Plates 129

4.4.2 Age Salience and Timing of Purchase by Status Level 131
4.4.3 Salience of Age and Demand for Premium Makes 137
4.5 Discussion and Conclusion 142
4.6 Supplementary Material 147
4.6.1 Additional Results 147
4.6.2 Theory 150

5 The Rule of Tome? Longer Novels are more likely to win Literary Awards 152
5.1 Introduction 153
5.2 Data 157
5.2.1 Literary Awards 157
5.2.2 Data Collection 158
5.3 Results 158
5.3.1 Summary Statistics 158
5.3.2 Logistic Regression 160
5.3.3 Relative Length on Shortlist 161
5.4 Potential Causes of the Bias 164
5.5 Conclusion 166
5.6 Supplementary Material 169

6 Conclusion 174

Bibliography 184

List of Figures

2.1 Response Screens for Judgment and Forced Choice Task 20
2.2 Judgment Task Distribution of Estimation Error and Mean Error by Condition (Experiment 1) 21
2.3 Judgment Task Distribution of Participant-Level Error (Experiment 1) 22
2.4 Average Error across Judgment Task (Experiment 1) 22
2.5 Performance on Forced Choice Task (Experiment 1) 25
2.6 Proportion correct by Condition and Trial Type (Experiment 1) 26
2.7 Judgment Task Distribution of Participant Level Error (Experiment 2) 32
2.8 Error across Judgment Task, Correlation to Prior (Experiment 2) . . . 35
2.9 Strategy Adjustment across Forced-Choice Task (Experiment 2) 40
2.10 Simulation of Forced Choice Results from Judgment Responses (Experiment 2) 42
2.11 Proportion of 'More' Response across Task (Experiment 1) 62
2.12 Forced Choice Task Strategy Adjustment (Experiment 2) 66
2.13 Household Bill Frame Introductory Screens (Experiment 1) 67
2.14 Slot Machine Bill Frame Introductory Screens (Experiment 1) 67
2.15 Forced Choice Task Instructions (Experiment 1) 68
2.16 Judgment Task Instructions (Experiment 1) 69
2.17 Additional Frames (Experiment 2) 69
2.18 Post-Task Questionnaire (Experiment 2) 70

3.1 Low Accessibility: Dispersed vs. Concentrated Fuel Cost Labels 87
3.2 High Accessibility: Dispersed vs. Concentrated Fuel Cost Labels . . . 87
3.3 WTP Input after Fuel Forecast 89

3.4 WTP Gap across maximum admissable Response Increases 96
3.5 Combinations Payment Schedule and Fuel Cost Frame 103
3.6 Introduction Page 103
3.7 Lease Contract Scenario Explainer 104
3.8 Payment Schedule 104
3.9 Personalised Running Cost Forecast 105
3.10 Response Screen to WTP 105
3.11 Distribution of Monthly WTP 106
3.12 Distribution of Annual WTP 106
3.13 Distribution of Three-Year (Deposit) WTP 106
3.14 WTP by Age Group 109
3.15 WTP Gaps below €600 109
3.16 Treatment Effects Across Thresholds of WTP 110
3.17 Illustration of Experiment Randomisation 110

4.1 Licence Plates in Great Britain and Ireland 123
4.2 Effect of Switching to Bi-Annual Plates in Great Britain (1999) 130
4.3 Effect of Switching to Bi-Annual Plates in Ireland (2013 131
4.4 Proportion of Sales in New-Reg Months by Car Make type and Country 132
4.5 Premium Make Market Share Trend: Great Britain vs. Ireland 138
4.6 Premium Make Market Share Trend: Northern Ireland vs. Ireland . . 139
4.7 Premium Make Market Share Trend: EU vs. Ireland 139
4.8 Inequality of Sales Pattern in Britain 147
4.9 Imports of Premium Makes and Exchange Rate 148
4.10 Imports and Total Registrations 148

5.1 Page Length of Winners and Losers 159
5.2 Relative Length on Shortlist 162

List of Tables

2.1 Judgment Task Mixed Effects Linear Regression Model (Experiment 1) 24
2.2 Forced Choice Task Mixed Effects Logistic Regression (Experiment 1) 27
2.3 Summary Statistics of Estimation Error by Frame (Experiment 2) . . . 33
2.4 Judgment Task Linear Mixed Model Regression Results (Experiment 2) 34
2.5 Forced Choice Task Proportion Correct by Condition (Experiment 2) . 36
2.6 Mixed Logit Regression Results (Experiment 2) 38
2.7 Priors for Sequence Sums (Experiment 2) 39
2.8 Judgment Task Mean Error by Estimation Strategy Type (Experiment 2) 41
2.9 Forced Choice Task Underestimation by Estimation Strategy Type (Experiment 2) 41
2.10 Sequence Effects and Estimation Error (Experiment 1) 56
2.11 Response Time Exclusion in Judgment Task (Experiment 1) 57
2.12 Mixed-Logit Results with Response Time Exclusions (Experiment 1) . 58
2.13 Mixed-Logit Results without Interaction Temrs (Experiment 1) 59
2.14 Forced Choice Task Mixed Effects Linear Probability Model (Experiment 1) 60
2.15 Frame-Specific Task Order Effects (Experiment 1) 61
2.16 Judgment Task Sequence Effects (Experiment 2) 62
2.17 Judgment Task Response Time Restrictions (Experiment 2) 63
2.18 Forced Choice Task Response Time Restrictions (Experiment 2) 64
2.19 Forced Choice Task Mixed Effects Linear Probability Model (Experiment 2) 65

3.1 Six Combinations of Payment Schedules and Fuel Cost Frame 85

3.2 Paired Combinations of Payment Schedule and Fuel Cost Frame . . . 86
3.3 Mean Annualised WTP for Payment Schedule - Cost Frame Combinations 90
3.4 Difference in WTP Between Combination Pairs 91
3.5 Random Intercept Regression Results 92
3.6 Random Intercept Regression Results for Combinations of Payment Schedule and Fuel Cost Frame 95
3.7 OLS Regression Results for Difference in WTP 107
3.8 Logit Regression Results WTP compared to Monetary Savings 108
3.9 Summary of Conditions and WTP 111

4.1 Timeline of Changes to British Licence Plate 125
4.2 Data Sources 127
4.3 OLS Regression Results Proportion of Make Annual sales in New-Reg Month 134
4.4 OLS Regression Results Make-type Monthly Sales Share 136
4.5 OLS Regression Results Premium Market Share (Make-Level) 141
4.6 OLS Regression Results Premium Market Share (Aggregated) 142
4.7 Tobit Regression Results Proportion of Make Sales in New-Reg Months 149
4.8 Missing Months of Sales Data for Northern Ireland 149

5.1 Summary Statistics Page Length of Winners and Losers 159
5.2 Logistic Regression of Award Outcome on Novel and Author Characteristics 160
5.3 Logistic Regression of Award Outcome on Relative Length on Shortlist 162
5.4 English-language Nobel Prize of Literature Winners since 1975 169
5.5 Linear Probability Model 170
5.6 Relative Length Regression 171
5.7 Number of Female Authors on Shortlist 172
5.8 Effect of Rank on Probability of Success 173

Chapter 1 Introduction

Behavioural economics brings psychological insights to bear on economic phenomena (Loewenstein, 1999). This thesis deals with how bounded-rationality shapes how individuals perceive and integrate numerical information in their decisions. The settings are diverse: Chapter 2 starts at the proverbial kitchen table where people are judging annual household bills. It moves to the car dealership in Chapter 3 and the decision of how much extra to pay for an improvement in vehicle fuel efficiency. Chapter 4 investigates a broader question in the same setting - how is demand affected by the format of an age identifier on the licence plate? Chapter 5 veers off the consumer choice track to peer inside the wood-panelled rooms where expert judges decide literary prizes. It analyses why the numeric attribute of novel length is apparently given extra weight in group decisions.

Despite their varied appearance, these essays have a unified purpose: to see "where the action is" - to borrow a phrase from Herbert Simon (1986) - in the determinants of economic choice. Simon demonstrated that assuming maximisation without careful empirical analysis of the utility function and *auxiliary assumptions* diminishes the scientific merit of economic explanations. In this sense, Chapters 2 and 3 investigate the auxiliary assumption that perceptions of numbers match their objective reality. Chapter 4 drops the rational-agent assumption of unlimited attention in signalling, in favour of limited attention. This rationalises the high demand for clear and obvious numeric signals of status. Finally, Chapter 5 investigates a high-stakes resource allocation problem and finds that the context of group decision making plausibly affects attribute-weights in multiattribute choice.

The subsections below provide a general introduction to each chapter, and include background context to motivate each investigation. Only a brief summary of method and results is provided because the emphasis is on contextualisation.

Preference or Perception?

"*People can only respond to incentives (the quality of goods on offer, their prices, and so on) to the extent that they are correctly perceived.*" – Woodford (2020)

The number line stretches all the way to infinity, and still Daniel Bernoulli found a way around it. In 1738, he resolved the St Petersburg paradox – why would people have a meagre willingness to pay for a gamble with infinite expected value? – by assuming utility to be a concave function of wealth. This insight was not overlooked, to put it mildly; expected utility theory became the cornerstone of microeconomics.

A little over 100 years later, Gustav Fechner published the *Elements of Psychophysics*, which charted the relationship between properties of the external world – temperature, light, time – and their internal sensations – heat, brightness, perceived duration. He found a common law in the relation between external stimulus and sensation: "in order that the intensity of a sensation may increase in arithmetical progression, the stimulus must increase in geometric progression". In other words, the mapping is concave. The natural perception of numerosity, called the "number sense" also follows this logarithmic pattern (Dehaene et al., 2007; Dehaene et al., 2008; Anobile et al., 2012). On the internal mental number line, small numbers are spaced well apart, but large ones are compressed together.

The irony might be obvious. Assuming a concave preference function over wealth distracted economists from the possibility that the representation of numbers, those raw inputs over which people 'optimise', might not match their objective size. This matters because genuine expected-utility valuations are considered beyond debate, or at least beyond the remit of economists.[1] But if choices reflect a cognitive bias instead, then inferring utility maximisation from observed choices is not valid.

However, just because people exhibit a compressed number line in psychology experiments, does not mean the bias would remain in a task where numbers have a

[1] as Stigler and Becker (1977) wrote, "*De Gustibus Non Est Disputandum*".

clear economic meaning and people understand the benefit of accurate perception; learning and high-stakes can eradicate biases (Levitt and List, 2007). They may also have a calculator or other decision aid at their disposal, and use that if adequately motivated and aware of the risk of error.

But using a calculator is more difficult when the pertinent numerical information is dispersed over time or place. In these situations people are likely to rely on their intuitive system to form an impression of the sum. The compressive internal number line predicts that intuitive summation will result in underestimation bias. In contrast, the implicit assumption in standard economic theory is that any errors should be random. This motivates the Chapter 1 research question: does underestimation bias persist in an economically familiar decision context?

We conduct two preregistered, incentive-compatible experiments that test the economic relevance of underestimation bias. The primary experimental manipulation is the framing of the number sequences to be summed. In the familiar frame, people sum monthly or bi-monthly household bills. In the abstract frame, they sum payouts from cartoon slot machines. The null hypothesis is that in both frames underestimation bias will manifest, and be of equal magnitude, because the compressive number line is invariant to framing. In the second experiment, two frames are added to dissociate factors that could attenuate underestimation bias according to signal detection theory, which models how people integrate contextual information to generate responses to stimuli. These additional frames isolate the effect of (i) people's priors for likely sums, and (ii) an asymmetric loss function for error. To probe the generalisability of underestimation bias, we elicit responses using a judgment task, where participants type their estimate for the sum, and also a forced-choice task, where participants compare their internal impression of the sum to a given number. Sequence length is varied in the first experiment only.

Results show consistent underestimation of approximately 6%. Underestimation occurs in both judgment and forced-choice tasks and is not diminished by imposing a

familiar framing, supporting the primary hypotheses. Moreover, clear learning effects imply the effect size likely marks a lower bound. These findings have implications for modelling the cognitive foundations of economic preferences. They also provide insight into how firms' pricing structures can exacerbate biases, causing economic loss.

Alternatives to Time Discounting

"Our telescopic faculty is defective . . . and we see future pleasures, as it were, on a diminished scale." – Pigou (1925)

Why do people appear to undervalue the monetary returns on energy-efficiency investments? This systematic undervaluation is called the energy paradox. The motivation in Chapter 3 is testing alternatives to time discounting for this behaviour. In his seminal research on high implied discount rates in energy-efficiency investments, Hausman referenced Pigou's lament about the defective telescopic faculty (Hausman, 1979, p. 51). Pigou attributed the defect to a failure of imagination; people could not bring the future into vivid focus, and hence choices were distorted toward present consumption. Modern conceptions of now-vs-later consumption decisions assume that the trade-off is perceived accurately (before preferences are applied). But underestimation bias indicates that people might undershoot the sum of the monetary savings. In other words, the fault in the telescope might be subtly different: the compressive number line might act like a warp in the lens, lowering the perceived position of 'total benefit' in the future sky.

In addition to underestimation bias, we test a second cognitive mechanism called concentration bias. Concentration bias is caused by limited attention. Attributes with larger differences attract attention. According to concentration bias, incremental savings on fuel costs are underweighted because the magnitude of fuel cost differences is much smaller than the magnitude of the upfront price difference. People focus their attention on the larger difference and this distorts the expected utility calculation.

Testing concentration bias in a controlled experiment was suggested by Allcott (2016) in a review of the energy-efficiency literature. This is the first study to take up that suggestion. The key point is that both underestimation bias and concentration bias can produce behaviour that appears identical to present-biased choices.

We employ a willingness-to-pay experiment to test these cognitive mechanisms against time discounting as explanations for the energy paradox. We elicit willingness-to-pay (WTP) for an improvement in fuel economy between two cars which are otherwise identical. The decision scenario is a 3-year lease contract which includes a deposit payment. The payment schedule for the investment is varied, as is the temporal framing of fuel costs for both cars. Participants view the fuel cost forecast and then input a WTP. Underestimation bias and concentration bias predict a pattern of results that is not matched by time discounting. Independently of this primary experimental manipulation, the presentation format on the forecast label is also varied to test an assumption of concentration bias, and a suggested mechanism underlying it. Results support the preregistered hypotheses, with the annualised WTP higher for more disaggregated payment schedules, and the WTP lower when the monetary savings are presented in a shorter temporal frame (e.g. monthly instead of annual). These findings indicate that the proportion of the energy efficiency gap attributed to present-biased preferences may be exaggerated. If the energy paradox is partly caused by a cognitive error, there is greater scope for policy intervention, relative to the scenario where choices genuinely reflect true preferences for present vs. future consumption.

Being Number One: Putting Status in Evidence

"*In order to gain and to hold the esteem of men, wealth must be put in evidence, for esteem is awarded only on evidence*" – Veblen, A Theory of the Leisure Class (1899)

People consume particular goods and services to signal a positive social image, also known as status-signalling. Standard models of conspicuous consumption assume

that status signals are received without error by their targets (also known as the signal receivers). This rationality assumption is central to the Spence (1973) signalling model. When signal receivers are rational, marks of wealth can be subtle and still the correct inference will be made. In contrast, the empirical evidence suggests people have limited attention. Moreover, signal receivers may not have any self-interest in attending to a status signal, as it could cause envy. The implication is that status signals may need to grab attention and have a clear meaning in order to be interpreted correctly. In others, to be worth buying, signals may need to be salient.

Chapter 4 investigates the value placed on clear and obvious status signals in the car market of Ireland and Great Britain. Both countries have numeric age identifiers on the licence plate. This acts as a status signal because new cars are more expensive than used ones. However, the age identifier is more salient on the Irish plate due to its location and its transparent format. The institutional history of licence plate regulations indicates that the cross-country variation in salience is plausibly exogenous.

The analysis shows that the difference in salience has a causal impact on two dimensions of demand for new cars: the timing of purchases and what type of cars are purchased. There is also an interaction between these dimensions. When the age identifier is more salient, a higher proportion of sales take place in the months when the new plate is released. The proportion of annual sales in these months is higher for car makes that are not conventionally considered status symbols. When the licence plate period shortens and the age identifier becomes more salient, there is a decline in the market share of conventional status signals (such as BMW). This implies substitution between status attributes.

The findings in Chapter 4 add to our understanding of how consumers value salient status signals under conditions of limited attention. At an applied level, they may inform labelling policies to nudge consumption patterns in directions that generate positive externalities, for example the optimal design of a 'green' licence plate. In

terms of theory, the results broadly support the salience models of economic choice (Bordalo et al., 2021), while also highlighting deficiencies in these models. In particular, the models are silent on how attribute *comprehension* influences decision weight. The obvious meaning of the age identifier on the Irish plate is likely a large part of its appeal, over and above its attention-grabbing properties.

When is Quantity Quality?

"*In decision making, more often seems better, yet in life, more is often not better*" (Hsee et al., 2005, p. 237).

Buying experience goods, also known as credence goods, can be risky because it is impossible to ascertain quality prior to purchase. Marks of assured quality can alleviate this risk. It is generally assumed that the experts who make quality judgments on behalf of the wider public do so in a unbiased manner.

Chapter 5 provides evidence of biased expert judgment. Judging panels for literary awards consistently favour longer novels on shortlists. The relationship is shown using all shortlisted novels for three prestigious awards - the Booker Prize, the Pulitzer Prize for Fiction, and the National Book Award for Fiction, covering a time span of 1963-2021. This finding is robust to controlling for author gender and Goodreads rating. A positive relationship between length and success probability is reasonable, but the size of the effect suggests other valid cues are underweighted in the process of selecting a winner.

This bias has many potential causes. One is the representativeness heuristic: longer novels resemble the tomes that constitute the foundations of the Western canon, and this may subconsciously sway judges. This bias could be exacerbated by the group context. Research on group decision-making shows that accountability to group members increases the use of normatively irrelevant cues that are not obviously irrelevant. Heightened sensitivity to magnitude under joint-evaluation is another potential cause (Hsee et al., 2005). When people are primed to evaluate using a

calculation mindset, magnitude receives greater weight than when evaluation is based on feeling (and does not need to be justified to others).

The results suggest an explanation for research showing that Booker Prize winners are not higher quality than shortlisted novels (Ginsburgh, 2003). The current findings have broader implications for the inferred quality of expert judgment and underscore the importance of using structured decision processes to minimise bias.

Conclusion

Chapter 6 concludes the thesis by summarising the findings, discussing limitations (both in terms of data and chosen method of investigation), policy implications, and possible future directions for research. This synthesis assesses the scientific value of the thesis to advancing our understanding of the cognitive underpinnings of economic decision-making.

Chapter 2 Consistent Underestimation in the Intuitive Summation of Monetary Amounts

Abstract

Many economic decisions rely on a fast, intuitive system of numerical cognition. When trying to judge the sum of a sequence of numbers, this system produces non-random errors: on average, it underestimates. This underestimation bias is thought to be caused by a compressive scaling of numbers when they are encoded internally. We present two preregistered, incentive-compatible experiments that tested the economic relevance of the underestimation bias. We varied the economic frame of sequences to be summed and deployed both judgment and forced-choice elicitation methods. Experiment 1 (n = 104) showed significant underestimation in the judgment task, with an overall mean bias of approximately -6%. Experiment 2 (n = 501) recorded persistent underestimation bias in both judgment and forced-choice task. Learning effects in the judgment task, and apparent strategy adjustment in the forced-choice task, imply the effect size likely marks a lower bound. These findings have implications for modelling the cognitive foundations of economic preferences. They also provide insight into how firms' pricing structures can exacerbate biases, causing economic loss.

This chapter is co-authored with Prof. Eleanor Denny and Prof. Pete Lunn, who collaborated on the design and gave helpful feedback on the analysis and write-up. These findings have been presented at the following international conferences: SJDM 2020 (joint 3rd place poster award), SABE (2021), SPUDM (2021), and has been accepted for presentation at IAREP 2022 (PhD travel grant winner) and the 2022 Max Planck Summer Institute on Bounded Rationality. I thank Benjamin Scheibehenne for providing experimental materials, and anonymous referees for comments that have improved the paper.

2.1 Introduction

Numerical information is integral to many economic decisions. Such information generally takes the form of prices, and other pertinent decision attributes may also be numeric. Standard economic models assume an accurate representation of number. But experimental evidence demonstrates that the 'number sense' is approximately logarithmic, obeying the Weber-Fechner law for the relationship between external stimulus and internal sensation (Fechner, 1948). The classic example of Fechner's Law is that doubling the luminence of a light less than doubles subjective experience of brightness. The numerical cognition literature shows the something similar happens with number perception: when numerical information is encoded internally on the 'mental number line', a compressive scaling is applied (Dehaene et al., 2007; Dehaene et al., 2008; Anobile et al., 2012).

But does this compressive scaling actually affect everyday economic decisions? If people apply rules of arithmetic when making decisions, it does not matter if their initial perceptions of number are compressed. Rules of arithmetic are straightforward to apply, but require effort. The compressed mental number line should not matter in practice, as long as people have sufficient motivation to slow down, deliberate, and apply what they presumably learned in school. However, empirical evidence indicates a slow, analytical approach is unlikely to be the norm. Even highly consequential decisions like choosing a car, mortgage, or retirement plan, tend to be made primarily based on intuitive judgment (e.g. Turrentine and Kurani, 2007; Lee and Hogarth, 2000; Benartzi and Thaler, 2007).

One implication of compressive scaling is that when a sequence of numbers is added intuitively, the sum is underestimated. This occurs because each number is compressed when encoded as a neural representation, then integrated to arrive at the sum, but - crucially - no correction of the initial compression occurs. This implies a systematic tendency for the intuited sum to undershoot the true total; in other words, an *underestimation* bias. People are more likely to use intuitive summation,

and hence be at risk of underestimation bias, in contexts where the way numerical information is presented stymies the use of analytical processes. For instance, when price information is temporally dispersed, like utility bills that arrive monthly or bi-monthly (every two months), or spatially dispersed, like prices in a supermarket. In these settings, a representation of total price will rely on intuitive summation as a default, unless there is foresight to use a decision aid like a calculator to record numbers as they appear.

Existing evidence for underestimation bias comes from an experiment using a student sample and an unincentivized field study (Scheibehenne, 2019). In the lab experiment, a sequence of numbers was shown rapidly on screen and participants typed their best guess for the sum. No economic context was provided, beyond describing the numbers as prices. In the field study, supermarket shoppers (n = 966) guessed the total price of their basket while in the checkout queue. The effect size was similar across both lab and field, with a median bias of approximately -5% (more details in Section 2.1.1). These results suggest the possibility that people routinely underestimate summations of monetary amounts. However, the supermarket differed from the typical meaningful financial setting in several ways, making any generalization tentative. It involved only low prices paid in a habitual context. It required consumers to judge how many items they had bought as well as how much each item cost. Responses may have been pulled down by outdated reference prices in memory (Mazumdar et al., 2005; Linzmajer et al., 2021), particularly when accuracy was not incentivized. Or there may have been a residual effect of marketing offers that made especially low prices salient.

Given little is known about the likely reach of underestimation bias, in this chapter we set out to test whether underestimation bias extends to familiar, substantive economic contexts. To do this, across two preregistered experiments, we vary three aspects of the decision: the contextual framing of numerical sequences, the method for measuring estimation accuracy, and the length of sequences to be summed. These manipulations provide insight into the scope of the bias. The motivation for each

manipulation is summarised below.

Embedding meaningful, familiar context within an experiment can improve performance (Alekseev et al., 2017). A familiar frame of obvious economic relevance could evoke the careful consideration and concern about inaccurate estimation a prudent consumer would display. This might counteract or even eradicate underestimation bias, a result that would imply the bias is unlikely to manifest in markets where consumers are similarly engaged and motivated. In Experiment 1, to induce high familiarity, we frame sequences as recurring household bills for services such as electricity, gas heating and internet. A control group is shown an unfamiliar frame, replicating the abstract number sequences that participants summed in Scheibehenne (2019). One possibility is that underestimation bias will be invariant to framing, because compressive scaling applies generally under conditions of intuitive summation. The alternative (preregistered) hypothesis is that the familiar framing will attenuate underestimation bias because the obvious economic consequence of the decision will motivate the participant to overcome the intuitive tendency to underestimate. In Experiment 2, we introduce two additional frames that isolate (i) aspects of associative price memory and (ii) a subjective loss function for estimation error, which in theory could explain variation in underestimation bias across contexts.

In addition to the framing manipulation, we make a methodological contribution by introducing a forced-choice elicitation method to test the accuracy of intuitive summation. In the forced-choice task, participants respond by choosing whether they think the true sequence sum is more or less than a number printed on screen. It is important to test whether underestimation bias extends to forced-choice, because in many economic contexts, choices are made by comparing internal representations of options. There is no requirement in these circumstances to explicitly generate and assign a number to each option. The standard judgment task in which participants type their estimate for the sum is also used. The issue with relying on typing tasks alone to make broader inferences is that internal representations of sums might be unbiased (on average), but become distorted in the process of number generation and

articulation. This interference has been documented: for instance, when asked to recall their salary, people's typed responses are biased by their level of job satisfaction, with high satisfaction pushing pushes responses upwards and the opposite for low satisfaction (Prati, 2017). It is unlikely internal representations of salary are equally biased. More generally, the preference reversal literature underscores the importance of ensuring that responses are not caused by the chosen elicitation method engaging specific psychological mechanisms (Cubitt et al., 2004; Tversky and Thaler, 1990). If underestimation bias is caused by a *general* psychological mechanism, it should be detected using both the forced-choice method and the standard judgment method.

Thirdly, how sequence length affects summation accuracy has not been tested before, despite this property being routinely varied in experiments that investigate intuitive averaging (e.g. Brezis et al., 2015; Tsetsos et al., 2012). We vary sequence length to be either six or twelve numbers long in Experiment 1. These lengths were chosen to match common monthly and bi-monthly billing frequencies. This manipulation has applied relevance because firms are increasingly using disaggregated pricing structures. We expand on this point and summarize related research in Section 2.1.2.

2.1.1 Numeric Cognition

Scheibehenne (2019) conducted both a laboratory experiment (n= 40) and a field experiment (n = 966) to investigate how people perceive and aggregate sequentially presented numerical information. In the lab experiment, forty university students (33 female) completed an estimation task with 75 trials. On each trial, 24 numbers in were shown rapidly in sequence. After each trial, participants typed in their best guess for the sequence sum. Sixty-five percent of all sequences were underestimated, with a mean bias of -5.5%, and nearly 90% of participants underestimating on average. The field study was conducted in two small supermarkets in Switzerland, where customers queuing at the check-out were invited to give an unincentivized, spontaneous estimate of the total value of their shopping basket. The median bias was -5%, 60% of baskets

were underestimated, and the magnitude of underestimation was increasing in the value of the basket of goods.

Compressive scaling was also the focus of research by Schley and Peters (2014). Their results showed that the precision of internal representations of numbers partly explained diminishing marginal utility for money. They showed this by having participants complete both a task where numbers were mapped to an unmarked number (Siegler and Opfer, 2003) and a hypothetical willingness-to-accept task. Moreover, the precision of symbolic mapping mediated the effect of numeracy, implying that higher-level numerical functions were founded on more basic or intuitive numerical cognitive processes. Schley and Peters (2014) also conducted a risky choice task. The linearity of the value function over gains and losses in this task was positively correlated with precision in the number line mapping task. More recently recently, Olschewski et al. (2021) showed that underestimation of the average of a sequence partly explains undervaluation of lotteries, which is typically attributed solely to risk aversion; a compressed mental number line had the best explanatory power for the findings.

Intuitive averaging has received more research attention than intuitive summation. Brezis et al. (2015) showed that an intuitive system for numerical judgment operates according to the same rules as visual perception, namely by aggregating the firing rate of specialised neurons. In this experiment there was a U-shaped relationship between intuitive averaging accuracy and sequence length. Accuracy decreased once working memory limits were exceeded, but increased with sequence length thereafter, because random noise in neuronal firing rates averaged out. Across all sequence lengths, participants tended to underestimate the sequence average. Tsetsos et al. (2012) investigated intuitive averaging using a forced-choice procedure and also recorded accuracy increasing with sequence length. To the best of our knowledge, this paper is the first to investigate this relationship for intuitive summation.

2.1.2 Household Finance and Bill Perceptions

Underestimation from intuitive summation is relevant to several confluences of the psychology and economics of decision-making. One is the low incidence of switching service provider. From the perspective of traditional economic cost-benefit analysis, consumer inertia implies high search costs and hassle costs of switching (Klemperer, 1995). Underestimating the total current outlay could be an additional source of inertia.

How sequence length affects estimation accuracy has applied relevance as digital services predominantly take the form of recurring payments (e.g. Spotify or Netflix). These services are a fast-growing component of the median consumption bundle. A shift to recurring payment has occurred in non-digital markets too (Gilbert and Zivin, 2014), in tandem with the growth in Automatic Bill Payment (ABS). These developments have been linked to higher consumption (Sexton, 2015) and higher debt levels (Fuentealba et al., 2021). The existing research emphasises an inattention channel as the underlying mechanism. However, if underestimation bias also plays a role, counteractive policy measures may need to address the pricing structure itself rather than only nudging consumers to attend to it.

Suggestive evidence on the effect of billing frequency comes from hypothetical willingness-to-pay (WTP) experiments, where more disaggregated payments usually increase WTP. This is often called the 'pennies-a-day' effect (Gourville, 1998). However, little is known about how billing frequency influences demand in the field, because selection effects (e.g. opting to pay small amounts regularly due to liquidity constraints) confound clean inference. Selection effects are avoided when firms change billing frequency for all customers. Wichman (2017) analyzed consumption responses when a water supplier switched neighbourhoods from bi-monthly to monthly billing in an essentially random order. Water consumption increased under monthly billing. The welfare analysis of this change assumed that more frequent billing increased the accuracy of price perceptions, and consequently that the greater

demand increased welfare. In other words, the implication was that the average household did not consume enough water under bi-monthly billing to satisfy its 'true' preferences. An alternative explanation for the consumption increase is that perceptions of total price were biased downwards by higher billing frequency. This would be consistent with the findings of a field experiment on savings (Hershfield et al., 2020): saving a small amount regularly was initially more popular, but also associated with a higher attrition rate, possibly because participants had mistakenly underestimated the total saving commitment.

2.2 Experiment 1

Intuitive processes of numerical cognition likely underlie estimation of the annual cost of recurring bills. In Experiment 1 we impose this process, and measure estimation accuracy under it. Experiments in the decision-from-experience paradigm (Hertwig, 2015) often present participants with a rapid sequence of numbers and then elicit a response. Scheibehenne (2019) used this paradigm to measure estimation error. We use the same method. In the judgment task, participants type in an estimate for the sequence sum. The incentive scheme for accuracy penalizes under and overestimation equally.

2.2.1 Method

Participants

One hundred four participants were recruited through the online platform Prolific. Participant quality is superior on this platform compared to Amazon mTurk, according to research (Gupta et al., 2021). The average age of participants was 30 (SD =10.3), 70 per cent were female, and 45 per cent had a degree at least. Participants were paid £2.25 to take part in the twenty-minute experiment and could win an extra £5 in each

task for accurate responses. Pre-screening ensured that all participants resided in the UK. To ensure these UK residents were familiar with the 'bills', the sequences were denoted in pound sterling (£) and attributed to popular UK service providers such as Virgin Media and Vodafone. The experiment was programmed in JavaScript using the Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Power calculations were performed to determine whether the intended sample size would be large enough to detect an effect similar in magnitude to that recorded in Scheibehenne (2019).[1]

Design and Hypotheses

The experiment employed a 2 (frame) x 2 (task order) between-subjects design. In the frame manipulation, one group saw sequences in the familiar household bill frame, the other group saw the unfamiliar slot machine frame (hereafter abstract frame). In the task order manipulation, one group did the judgment task first, the other did the forced-choice first. All participants completed both tasks. Each task contained two practice trials and 24 incentivized trials, meaning 5,408 observations were recorded in total (4,992 with practice trials omitted).

In the bill frame, the sequences were described as bills for electricity, gas, car insurance, health insurance, TV, internet, phone, and gym membership. In the slot machine frame, they were described as pay-outs from cartoon slot machines, labelled A-H. In both tasks, twelve monthly (length = 12) and bi-monthly (length = 6) sequences were interleaved in a random order. A binary split was imposed on the sequence sums. Sums larger than the median were always associated with electricity, gas, health insurance and car insurance in the bill frame (and slot machines A-D in the slot machine frame). The smaller sums were associated with TV, internet, phone and gym membership in the bill frame (and slot machines E-H in the slot machine

[1] For the judgment task, power calculations showed 176 independent observations per group were required to detect with 80% power an effect size of 0.3 (Cohen's d) in a difference-in-means test with alpha set at 0.05. The design planned for 1200 observations (24 observations per participant) which provided sufficient power. A separate power test was not performed for the forced-choice task. However this was conducted for Experiment 2 (see Chapter Supplementary Section 2.5.1).

frame).

The sum and variance of the sequences was generated in a way that avoided a confound with the sequence length manipulation. The final range of sums was an approximately uniform range from £222-£958, with a mean of £547 and a median of £549. See Chapter Supplementary Material for details of how the sequences were generated.

In the forced-choice task, participants decide whether the true sum is greater or smaller than a given referent. On trials where true sum is greater, the correct response is to answer 'More'. These trials are denoted as 'More Correct' trials, or **MC** for short. Similarly, the abbreviation **LC** denotes 'Less Correct' trials, where the true sum is less than the referent. Trials were evenly split between MC and LC and their order randomly dispersed at the individual level. Trial difficulty was determined by the percentage difference between the correct sum and referent. This percentage difference is called the increment: it becomes more difficult to select the correct response as the increment gets smaller. The increment levels were set to 0.03, 0.09, and 0.15 of the correct sum, and balanced to be positive or negative. Positive increments correspond to LC trials, and negative increments to MC trials. As an example of an MC trial, consider a true sum of £500 and an increment of -0.09. This equates to a monetary increment of £45, producing a referent of £455 (i.e. £500 - £45). All participants made four responses at each of the six increment levels. Recall the prediction that participants will underestimate sequence sums on average. This implies a higher proportion of LC trials should be answered correctly than MC trials.

Summary of Hypotheses

H1a: In the judgment task, underestimation will be recorded in both frames.

H1b: Regarding relative underestimation in the judgment task, the null hypothesis is there will be no difference between frames. The alternative hypothesis is that

underestimation will be attenuated in the bill frame.

H2: In the forced-choice task, we hypothesis that underestimation will be recorded in both frames. Regarding relative underestimation, the pattern in H1b applies.

H3: The magnitude of underestimation will increase with sequence length.

Procedure

Participants first consented to abide by the experiment rules, which included not using pen and paper or a calculator. Participants were told they would see sequences of six or twelve numbers, with each number on screen for only half a second, and their job was to estimate the sum as best they could. These sequences were described as either 'bills' or 'slot machine payouts', depending on condition (see Experimental Instructions in Chapter Supplementary Material for full details).

After completing two practice trials, the lottery scheme to incentivize accurate responses was explained. The incentive scheme for the judgment task worked as follows: if a participant's absolute accuracy was in the top half of the sample on a given trial (e.g. if the median accuracy was 87% and her response was 90% accurate) her ID would be entered into a draw to win £5. Consistently being in the top half of the sample meant 24 entries to the lottery. Five IDs would be drawn from the pool of entries and each would win £5. This is a type of winner-take-all lottery (Cason et al., 2020) except with multiple winners. The incentive scheme was slightly different for the forced-choice task. Participants were told that five IDs would be drawn randomly, and then a trial number would be selected at random. Each chosen participant who had answered the selected trial correctly would win £5. These incentive schemes aimed to strike a balance between incentivizing effort and engagement without making cheating too tempting (which could happen under a deterministic tournament incentive where the best performer(s) win).

In the judgment task, participants received feedback about their average absolute

accuracy after every eight trials. The wording and form of this feedback was taken from the instructions of Scheibehenne (2019). This feedback did not give an indication of whether participants were over- or underestimating on average.

Each trial began with an introduction screen that displayed the type of bill (or slot machine) that was about to be shown and the length of the sequence, e.g. '12 monthly bills for electricity' (or '12 monthly payouts from slot machine E'). Participants clicked a button to begin the trial. Each number was shown on screen for 500 milliseconds and followed by a fixation cross (200 milliseconds). The response screen appeared after the final fixation cross.

As shown in Figure 2.1, the judgment task response screen took the form of an input box that appeared with a prompt. In the forced-choice task, the response screen displayed the referent and instructed the participant to indicate whether they thought the true sum was 'More' or 'Less' than the referent.

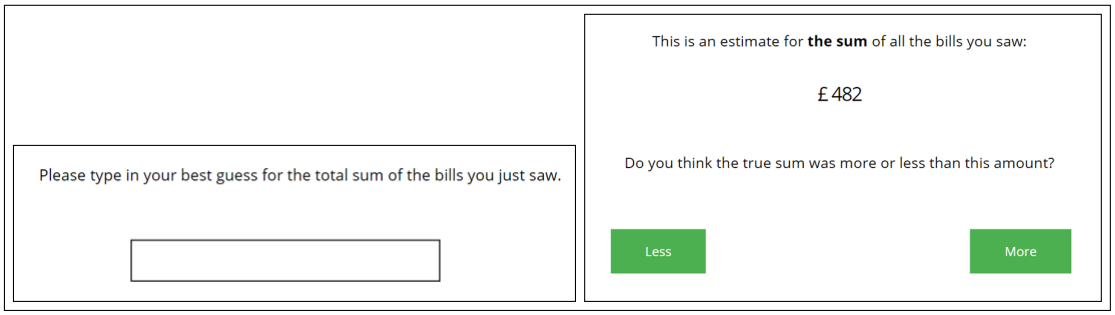

Figure 2.1: Response screen for judgment task (left panel) and forced-choice task (right panel).

2.2.2 Experiment 1 Results

Judgment Task Results

The primary dependent variable in the judgment task is the percentage error in response: (estimate minus true sum)/ true sum. The overall mean estimation error

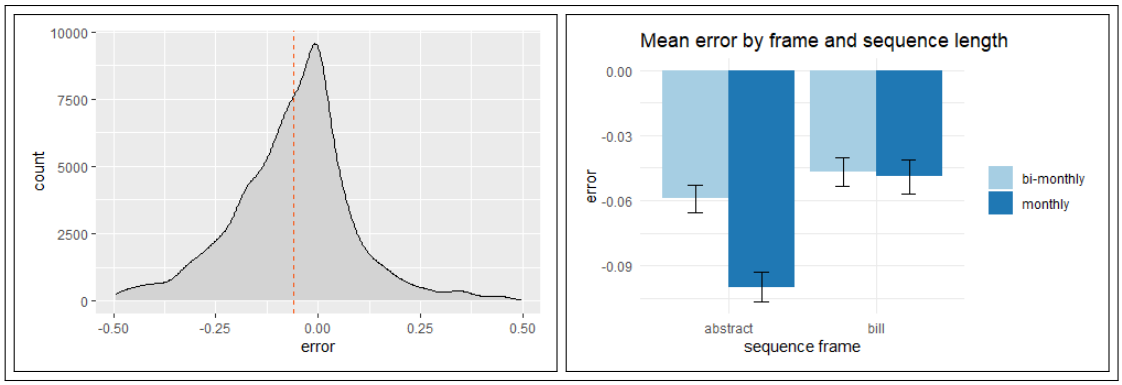

Figure 2.2: Distribution of estimation error (left). Mean error for four frame-length combinations (right).

is -0.064 (SD = 0.176) and the median error is -0.05. Figure 2.2 (left panel) shows the distribution of estimation error. In total, participants underestimated 67% of sequence sums. Figure 2.2 (right panel) shows the mean estimation error by frame and sequence length. In the bill frame, the mean error for shorter sequences is -0.047 (SD = 0.166) and for longer ones is -0.049 (SD = 0.197). In the abstract slot machine frame however, there is a clear difference by length: the mean error for the shorter bi-monthly sequences is -0.059 (SD = 0.159) and for longer monthly sequences is -0.099 (SD = 0.17). Pooling across sequence length, the slot machine frame records significantly greater underestimation than the bill frame (two-sample t-test, $p < 0.0001$). Pooling across frames, underestimation is significantly greater for longer sequences compared to shorter ones (two-sample t-test, $p = 0.0023$). Eighty-six of the 104 participants underestimated the sequence sums on average, 14 overestimated, and four participants' average error was less than half a percentage point, which we round to 'no error'. Ten of the 14 overestimators were bill frame participants.

One quarter of participants recorded a median estimation error of at least -0.095. Figure 2.3 above shows the distribution. We also measure the extent to which estimation accuracy improved over the course of the judgment task. The mean error by trial number is shown in Figure 2.4 below. The trend indicates clear improvement that tapered off after about six trials.

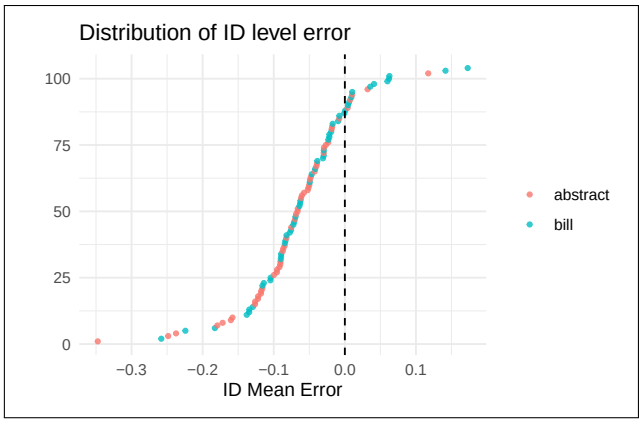

Figure 2.3: Distribution of mean error at the ID level for abstract frame (orange dots) and bill frame (green dots). Black dashed line marks zero-error point.

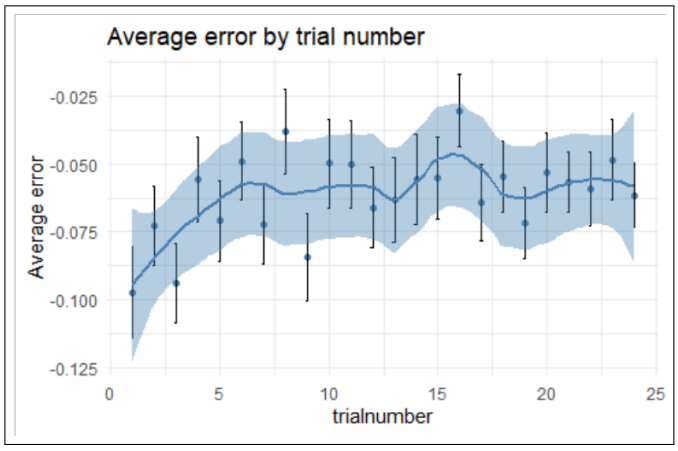

Figure 2.4: Average error by trial number in judgment task.

Regression Models

We conduct mixed-effects linear regression analysis with a random-intercept at the participant level to account for the repeated-measures in the data. The percentage error is the dependent variable. The covariates are dummy variables for frame (0 = slot machine, 1 = bill) and sequence length (0 = bi-monthly, 1 = monthly) and the standardized sequence sum. We omit responses that scored less than 50% on absolute accuracy, and also omit responses from two participants who indicated in the post-task strategy question that they used calculation aids. This leaves 2401 responses out of the full sample of 2496. (See Table 2.11, Column 1 in Chapter Supplementary

Material for regression output with entire sample.)

Column 1 of Table 2.1 shows that the coefficient on the bill frame dummy is not significant (β = .0218, p = 0.11). The coefficient on the sequence length dummy ('Longer') is negative and significant, indicating stronger underestimation for longer sequences (β = -.0143, p = .009). The coefficient on standardized sum is also negative and highly significant, indicating stronger underestimation for larger sums (β = -.014, p < .00001). The constant in the model is -0.0657. This gives an estimate of mean error when all other variables are zero. Given how the dummy variables were coded, this constant can be interpreted as the mean underestimation for a bi-monthly abstract sequence that summed to approximately £550.

Column 2 reports results from a specification where we interact the frame and sequence length dummy variables. The interaction is positive and highly significant (β = .0302, p = .006), as suggested by the bar chart in Figure 2.2. The sequence length dummy becomes more negative (β = -.029, p < .0001). The coefficient on the standardized sum is unchanged. Adding demographic covariates (age, gender, self-reported mental maths ability on a 1-7 scale) has no effect on the results (unreported analysis).

Column 3 adds a dummy variable for task order. Completing the forced-choice task first reduces underestimation in the judgment task (β_{order} = 0.0264, p = 0.048). Summary statistics, which are presented in the Chapter Supplementary Material, indicate that this order effect is present for both frames: for the bill frame, the order effect reduces mean underestimation from 6.2% to 3.5%. For the abstract frame, it reduces underestimation from 9.1% to 6.8%. The most likely explanation is that the referents in the forced-choice task calibrated the mental number line (Dehaene et al., 2008), reducing the scale of subsequent underestimation.

While including the standardized sum was an indirect test of Steven's power law, we also test it directly. We do this by estimating the power function. To do this, we regress the log of the response on the log of the true sum. This is mathematically equivalent

	Simple	Interaction	Order
Bill frame	0.0218	0.00684	0.00578
	(0.0136)	(0.0147)	(0.0144)
Longer	-0.0143**	-0.0291***	-0.0291***
	(0.00548)	(0.00766)	(0.00766)
Standardized Sum	-0.0143***	-0.0143***	-0.0142***
	(0.00271)	(0.00270)	(0.00270)
Billframe*Longer		0.0302**	0.0302**
		(0.0109)	(0.0109)
Choice Task First			0.0264*
			(0.0134)
Constant	-0.0657***	-0.0584***	-0.0716***
	(0.00991)	(0.0103)	(0.0121)
Observations	2401	2401	2401

Standard errors in parentheses
Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.1: Mixed-effects linear regression for Experiment 1 judgment task. Dependent variable is percentage error in response.

to estimating $y = \alpha * x^{\beta}$, where α is a scaling factor and β is the exponent that determines the degree of curvature. Compression implies a β coefficient of less than 1. The estimated β coefficient is 0.96 ($p < 0.0001$, 95% confidence interval = [0.945 - 0.976]) and the estimated scaling coefficient is 0.169 ($p = 0.001$, 95% confidence interval = [0.07 - 0.266]). Consistent with the earlier results, when dummy variables are added for frame and trial type, compression is markedly more pronounced for longer sequences in the abstract frame.

Forced-Choice Results

The level of performance is shown in Figure 2.5. The mean and median score was 16 out of 24. The minimum score was 8 and the maximum was 22. Eleven participants performed at worse than chance levels. These are shown to the left of the red dashed line in Figure 2.5.

The proportion correct is consistently higher on LC trials. The differential is increasing

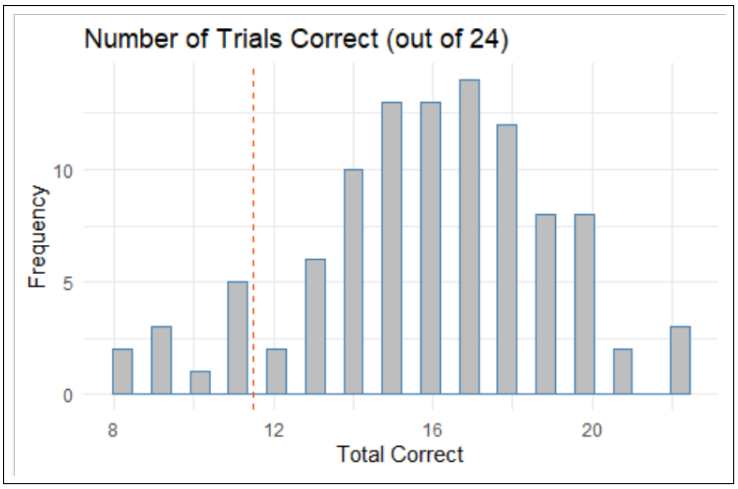

Figure 2.5: Histogram of number of correct responses (out of a possible 24) in forced-choice task.

in the level of trial difficulty: on 'easy' trials with a 0.15 increment, 75% of LC trials are correct, compared to 72% of MC trials; on medium trials with a 0.09 increment, the figures are 71% and 67%; on the most difficult trials with a 0.03 increment, 60% of LC trials are correct compared to 55% of MC trials. Overall, the proportion correct is 69% for LC trials and 64.6% for MC trials. This difference is significant at the five percent level (two-sample test of proportions, p = 0.0108, one-tailed). However, breaking the results down by frame uncovers significant underestimation in the abstract frame (two-sample test of proportions, p=0.0013) but no evidence of underestimation in the bill frame. Moreover, absolute accuracy is not better in bill frame, where 65.5% of trials were answered correctly, compared to 68.2% in the abstract frame. This difference falls short of significance however (p = 0.15).

Regression Model

We conduct a random-intercept logistic regression to account for the repeated measures in the data. The dependent variable is a binary indicator for whether the response was correct. Covariates are dummy variables for trial type (0 = LC, 1 = MC), for frame, and for sequence length, and a categorical variable for trial increment (the reference case being the easiest trials with a 0.15 increment). The trial-type dummy

is separately interacted with frame and sequence length dummies (See Table 2.13 in Chapter Supplementary Material for regression output without interaction terms, and Table 2.14 for Linear Probability Model results). Task order and trial number are also included to account for any learning effects.

Results show that on LC trials, the log-odds of a correct response are significantly lower in the bill frame (β = -0.34, p = 0.026). The interaction between MC and bill frame is positive and significant (β = 0.415, p = 0.019). Longer sequences are significantly less likely to be answered correctly on MC trials (β = -.545, p = 0.002). The log-odds of a correct answer improve significantly as the task progresses, but the effect size is small (β = 0.0156, p = 0.015). Column 2 restricts the sample to those who answered at chance levels or better (i.e. 12 correct out of 24). The pattern of results is essentially unchanged, with one exception: in the restricted sample, the order effect is more pronounced (β = 0.256, p = 0.009). The numerous interactions between frame, sequence length and correct response may be difficult to interpret numerically. Figure 2.6 below plots the proportion correct by trial type and sequence length for each frame (increment levels have been pooled).

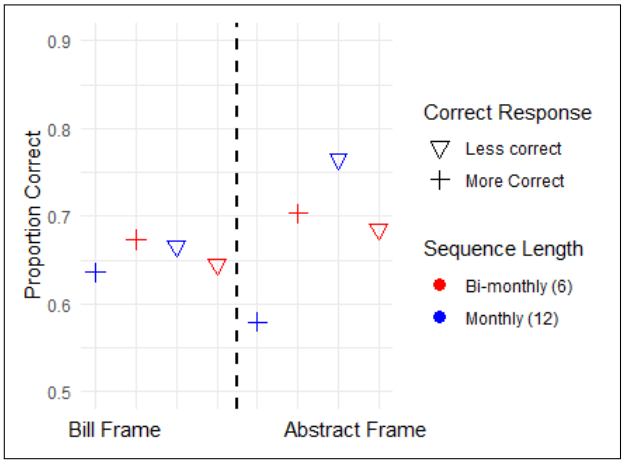

Figure 2.6: Forced Choice Task Proportion correct by sequence length and trial type for bill frame (left of dashed line) and abstract slot machine frame (right). Each data point is the average of 312 binary responses.

	Full Sample	Restricted
MC trial (ref. = LC)	-0.141	-0.140
	(0.155)	(0.166)
Bill frame	-0.340*	-0.337*
	(0.153)	(0.139)
MC*Bill frame	0.415*	0.393*
	(0.177)	(0.191)
0.09 Increment (ref = 0.15)	-0.218	-0.206
	(0.112)	(0.122)
0.03 Increment	-0.747***	-0.849***
	(0.109)	(0.117)
Longer	0.209	0.196
	(0.128)	(0.138)
MC*Longer	-0.545**	-0.520**
	(0.179)	(0.192)
Forced-Choice Task First	0.200	0.256**
	(0.122)	(0.0985)
Trial Number	0.0156*	0.0157*
	(0.00638)	(0.00688)
Constant	0.959***	1.075***
	(0.174)	(0.172)
Observations	2496	2184

Table 2.2: Experiment 1 Forced-Choice Task Mixed-Effects Logistic Regression.

There is a notable lack of variation in the proportion correct across trial types in the bill frame (left of dashed line). In contrast, in the abstract frame the proportion correct varies substantially. On LC trials, where underestimation bias was advantageous, performance is better on longer sequences than shorter ones (two-sample test of proportions, p = 0.0124, one-tailed). To see this in the Figure, note that in the right panel, the blue triangle lies above the red triangle. In contrast, on MC trials, where underestimation bias is a disadvantage, performance is better on shorter sequences (two-sample test of proportions, p = 0.0007, one-tailed). Note that the red cross is considerably higher than the blue cross in the abstract frame panel. Both of these directional effects are in line with the hypothesis that underestimation would be increasing in sequence length.

Lastly, regarding task order effects, there is no indication of a performance boost

from doing the judgment task first. In fact, there is evidence of a negative order effect for the bill frame, but only on LC trials. Bill frame participants who completed the forced-choice task first answered 71% of LC trials correctly. This figure falls to 61% for those who had already completed the judgment task. Without this order effect, the bill frame may have recorded significant underestimation too (see Chapter Supplementary Material for details).

2.2.3 Experiment 1 Discussion

Underestimation was clearly evident in both frames in the judgment task, including an everyday context of accumulating household bills. However, the interaction between frame and sequence length, where underestimation was greater for longer sequences - but *only* in the abstract frame - was not hypothesized. Why did this occur?

One possibility is that the familiar bill frame induced a strategy adjustment, because it brought to mind a contextual factor that informed responses. This contextual factor may have acted as a countervailing force to the stronger underestimation that occurred in the abstract frame longer sequences. Contextual factors are integrated in what signal detection theory calls the "decoding stage" of decision making (Green et al., 1966). The two candidate factors we considered are: (i) building in the relative cost of error in each possible direction and (ii) incorporating a prior for the likely total bill.

Regarding (i), the technical term for situations where the costs of error is not the same in both directions is an 'asymmetric loss function'. If participants thought that underestimating bill totals was 'worse' than overestimating, they may have held an asymmetric subjective loss function for estimation error. This is not entirely arbitrary: overestimation of bill totals can reasonably be perceived as application of a precautionary principle, like erring on the side of arriving too early for a flight rather than too late.

Alternatively, incorporating priors which on average were larger than the true sequence sums, may have counteracted the tendency to underestimate. But why would this not also apply to the shorter sequences? Both contextual factors may have received greater weight when estimating longer sequences: with working memory limits exceeded, participants who were highly uncertain over the sum would reasonably incorporate other information to guide their response.

However, the effects of these two contextual factors were confounded in the first experiment. The familiarity of the bills meant people likely had priors for their totals. And the fact they were bills meant an asymmetric loss function could apply. The second experiment isolated these factors by introducing new conditions which separated their effects. A second experiment was also necessary to test the finding of underestimation in the forced-choice task, which was suggestive but far from definitive.

2.3 Experiment 2

Experiment 2 used the same method as Experiment 1, except with a simplified design and the addition of two new frames. We designed these frames so that the two contextual factors, priors and the asymmetric loss function, would pull estimates in opposite directions in each frame. This allowed the effect of each factor to be isolated.

The new frames were a 'non-familiar bill' frame and a 'familiar non-bill' frame. For the non-familiar bill frame, we framed the sequences as bills for the cost of feeding exotic pets (reptiles, birds of paradise, zebras, etc.). Most people have little experience of these feeding costs. Using priors for feeding common household pets would almost certainly pull estimates downwards. Therefore, we hypothesized that *attenuated underestimation* in this frame, relative to the slot machine frame, could only be attributed to an asymmetric loss function induced by bills in general, where it is better to err on the side of caution and overestimate. We presented a montage of

cartoon animals but did not refer to specific ones by name. Instead, the bills were for feeding exotic pets labelled A-H.

For the 'familiar non-bill frame', we framed the sequences as savings towards a short vacation.[2] The concept of saving for a once-off expense is a familiar one. And for experiential goods in particular, people prefer to pay upfront than on credit (Quispe-Torreblanca et al., 2019). Savings is also a domain where the precautionary principle, if it applies, flips towards underestimation being less subjectively costly. This means any asymmetric loss function should strengthen underestimation. Therefore, we hypothesized that *attenuated underestimation* in this frame compared to the slot machine condition could be attributed only to priors for the cost of the vacation being higher than the sums displayed, dragging responses upwards. The eight types of activity vacation we used were hiking, camping at a music festival, surfing, scuba diving, playing golf, skiing, sailing, and a river cruise.

2.3.1 Method

Participants

Five-hundred-and-one participants were recruited from the website Prolific and were paid £1.40 to take part in the experiment which took 12-15 minutes to complete. The sample was broadly representative: 56% were female and 44% male. The average age was 39 (SD = 14). Fifty-four percent had a degree or higher. Participants were randomly allocated to one of the eight conditions described below.

[2]Because participants were based in the U.K, we used 'holiday' instead of vacation in the experimental instructions.

Design

We employed a 2 (task) x 4 (frame) between-subjects design. Sequence length was fixed at 12 on all trials. The task length was 24 in the judgment task, as in Experiment 1, but extended to 32 trials for the forced-choice task. The increments were lowered to 0.09, 0.06 and 0.03 in order to increase the statistical power to detect an effect. Half of all increments were set at 0.06, with one quarter at 0.09 and 0.03 respectively. There was an even split between positive and negative increments.

We aimed to split the sample 3:2 between forced-choice and judgment, because the binary-response nature of the forced-choice task means more observations are required to obtain equivalent statistical power. The decision not to split the sample evenly between the tasks was taken during the process of computing the required power. We aimed for 9,600 observations in the choice task (10,200 including practice trials) and 4,800 responses in the judgment task (5,200 including practice trials) meaning the final dataset would comprise over 15,000 responses. These figures were determined based on power analyses that took a very conservative approach to how the intraclass correlation between an individual's responses would affect the standard errors for inferential statistical tests.[3] Full details are described in the Chapter Supplementary Material Section 2.5.1.

Immediately after completing the task, participants were asked to type out any estimation strategies they used. Participants (except in the slot machine frame) were also asked whether they considered real-life costs when deciding on their response. Regardless of response to this question, a follow-up asked whether they thought the *real-life costs* for the type of product they were shown (household services/feeding exotic pets/short vacation) were higher, lower, or about the same cost as the monetary sums in the task. Household bill frame participants were also asked to estimate the

[3]The power calculations included the intraclass correlation from Experiment 1. This addition inflated the required sample sizes. For the judgment task, 810 observations were required per group for 80% power given a 5% alpha level and an effect size of d = 0.3. For the forced-choice task, to meet the same level of power, 681 observations per group were required.

average annual household expenditure for three common bills – electricity, gas, and mobile phone. They were told to disregard the sequences they saw previously before making these estimates. These questions were included to test whether contextual factors had an effect on responses.

Materials and Procedure were the same as in Experiment 1. The Experimental Instructions in Section 2.5.5 provides screenshots of the experimental interface.

2.3.2 Experiment 2 Results

Judgment Task

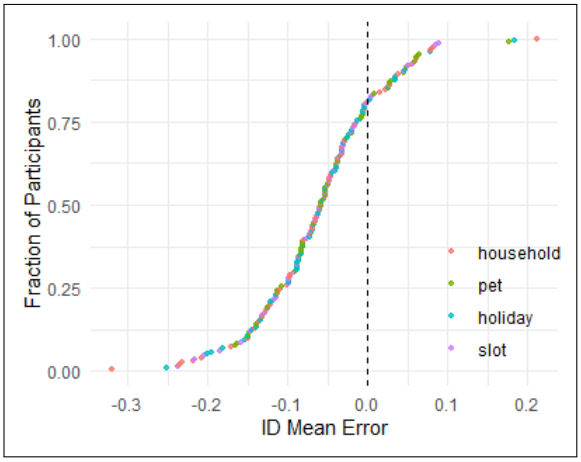

Figure 2.7: Distribution of ID-level error in Experiment 2 judgment task.

The judgment task was completed by 198 participants. Results show persistent underestimation. In the full sample of 4752 responses, the median error is -0.057 and the interquartile range of error is -0.16 to 0.02. Sixty-eight percent of sequences were underestimated. The distribution of participant level mean error is shown in Figure 2.7 below. The sigmoid pattern resembles the distribution from Experiment 1 (Figure 2.3).

A preregistered exclusion criterion was that responses below 50% absolute accuracy

would be omitted. This criterion was designed to prevent mistakes such as adding an extra zero, or inputting averages instead of sums, from adding noise to the data. The main result is not sensitive to this exclusion. In the restricted sample, the median error is -0.055 and the mean error is -0.06. Summary statistics for each frame are shown in Table 2.3 below.

Frame	Mean	S.D	Median	25th PCT	75th PCT	Obs. (IDs)
Household Bills	-0.067	0.16	-0.061	-0.16	0.02	1045 (48)
Exotic Pet Bills	-0.049	0.143	-0.048	-0.14	0.03	1140 (48)
Vacation Savings	-0.06	0.14	-0.058	-0.14	0.01	1228 (52)
Slot Machine	-0.062	0.139	-0.055	-0.14	0.01	1147 (50)

Table 2.3: Summary statistics of estimation error by frame.

Underestimation was strongest in the household bills frame (-6.7%) and weakest in the exotic pets frame (-4.9%). This difference is significant (two-sample t-test, p = 0.007). However, it is not significant when we conduct mixed-effects linear regression analysis (β = .016 p = 0.33). The reference case is the household bills frame.[4] As in Experiment 1, the standardized total sum is a significant predictor of estimation error, with stronger underestimation for larger sums (β = -0.0132, p < 0.0001). This is in line with Steven's power law for subjective perceptions of intensity.

As in Experiment 1, we conducted a log-log regression model to estimate the degree of compression directly. Again, we found an estimated β coefficient of 0.96 for the exponent. But due to the larger sample (n = 4560), the 95% confidence interval is narrower: [0.951-0.971]. The estimated scaling coefficient is 0.17, with a 95% confidence interval of [0.10-0.23].

[4]See Chapter Supplementary Materials for sensitivity checks to omitting responses at the tails of the response time distribution.

	(1)	(2)	(3)
Bills for Feeding Exotic Pets	0.0161	0.0121	0.063
	(0.0165)	(0.0159)	(0.0479)
Vacation Savings	0.00459	-0.000146	0.042
	(0.0162)	(0.0157)	(0.047)
Slot Machine	0.00314	-0.000841	0.079
	(0.0165)	(0.0159)	(0.047)
Standardized Sum	-0.0132***	-0.0132***	-0.0202***
	(0.00180)	(0.00180)	(0.0022)
Constant	-0.0647***	-0.0608***	-0.106***
	(0.0119)	(0.0117)	(0.0339)
Observations	4560	4445	4736

Table 2.4: Experiment 2 Judgment Task Regression Table. Column 2 specification drops participants who had five or more responses that scored less than 60 percent on absolute accuracy, which was another preregistered exclusion criterion. Column 3 includes all responses between €1-1999.

Learning and Influence of Priors

As in Experiment 1, we find evidence of a reduction in underestimation across the judgment task. This is shown in Figure 2.8 (left panel). Practice trials are included to the left of the red dashed line. This improvement in performance may constitute learning by participants. They were given feedback on absolute accuracy after every eighth trial, a feature which was included in order to replicate Scheibehenne (2019) as closely as possible, and may have correctly guessed the most common direction of error and adjusted accordingly. Learning is rapid initially, but tapers off well below the correct-on-average threshold. The trend is less noisy than for Experiment 1, which was shown in earlier in Figure 2.4, probably due to the larger sample size and the absence of task-order effects.

In the household bill frame, after completing the estimation task, participants were asked to guess the actual average annual UK household bill for electricity, gas, and the individual bill for a mobile phone. They were told to disregard the sequences for these bills they saw in the task. The responses for the three bills were averaged.

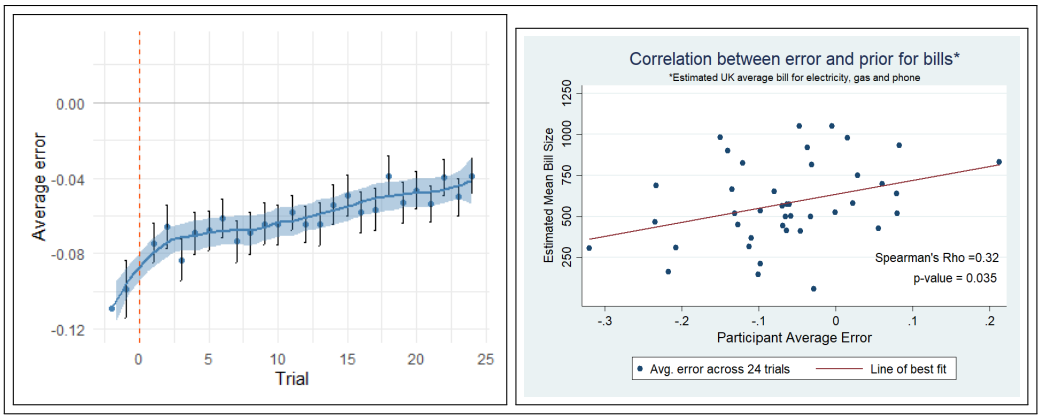

Figure 2.8: Left panel: Mean estimation error across 24 trials (and practice trials to the left of red dashed line). Right panel: Correlation between mean error and prior for bill costs.

We calculated the correlation between this average and participant's mean level of underestimation to measure the effect of priors. Higher priors should attenuate underestimation. The correlation is positive and significant, with a Spearman's rho of 0.33 (p-value = 0.035). The relationship is shown in Figure 2.8 (right panel). The case for making a *causal* interpretation of the relationship in Figure 2.8 would be strengthened if there was *no correlation* between real-bill estimates and mean error for participants in the other frames. However, we did not ask participants in other frames to estimate real annual bills because we reasoned it might be a jarring transition from estimating vacation savings or slot machine payouts. We acknowledge this limitation. Nevertheless, the forced-choice results below provide some support for a causal interpretation.

2.3.3 Forced-Choice Results

The 303 participants in the forced-choice task each made 32 responses, making 9,696 observations in total. For LC trials, the overall proportion correct was 0.66; for MC trials, the proportion was 0.614. This difference is statistically significant at the one-percent level (two sample test of proportions, $p < 0.0001$). This is clear evidence

of underestimation bias. The worse performance relative to Experiment 1, where 69% of LC trials and 64.6% of MC trials were answered correctly, is due to the smaller increments which made the task more difficult.

At the individual level, the median score was 21 out of a maximum 32, and the mean score was 20.4 (SD = 4.2). Twenty-eight participants performed at below-chance levels (less than 16 correct). Closer inspection revealed that the low level among this group was not random. Instead, the worse-than-chance performers answered a higher proportion of MC trials correctly (two-sample test of proportions, p = 0.0007). This is consistent with exhibiting underestimation and erroneously doing the task backwards - in other words, asking oneself "*is this referent more or less than the true total?*" These mistaken performers were not spread evenly across the four frames. Instead, they were concentrated mostly in the vacation savings and exotic pet bills frames (12 and 9 participants out of 28 respectively). For this reason, in order to make meaningful comparisons between frames, we omit participants who performed worse than chance from the subsequent analysis. This leaves 8,800 observations from 275 participants. In this subsample, the proportion of less trials answered correctly is exactly 69.5% and the proportion of more trials correct is 63.2%. Table 2.5 below shows the proportion of LC and MC trials answered correctly for each of the four frames. Underestimation bias is strongest in the household bills frame (p <0.0001).

	Less Correct (LC)	More Correct (MC)	Gap	Avg. Accuracy
Household Bills	.71	.59	.12	.65
Feeding Exotic Pets	.71	.66	.05	.685
Vacation Savings	.68	.66	.02	.67
Slot Machines	.68	.63	.05	.655

Table 2.5: Proportion correct on LC and MC trials by frame for 275 participants who performed at chance or better. Proportions have been rounded.

It is highly significant for the slot machine frame (p = 0.0026) and significant at the five-percent level in the exotic pet bills frame (p = 0.0108). In the vacation savings frame, the two-percentage point difference in performance between LC and MC trials

is not significant (p = 0.125). However, this frame also has the smallest number of observations, due to a higher proportion of worse-than-chance performers. We return to this issue after we present the results of the mixed-effects logistic regression models.

Regression Model

As in Experiment 1, the dependent variable is a binary indicator for correct response. Covariates are a dummy for trial type (0 = LC, 1 = MC), a categorical variable for frame (the household bills frame being the reference case), an interaction of the trial type dummy with frame categories, and also indicators for the trial increment (reference case: 0.09 increment). Table 2.6 below presents the results. We focus on Column 2.

As suggested by the descriptive statistics, a correct response was less likely on MC trials (β = -0.522, p < 0.0001). The effect size is similar in magnitude to the reduction in log-odds of a correct response when the increment decreases from 0.09 to 0.03 (β = -0.603, p < 0.0001). For LC trials, there is no difference in precision by frame. However, the interaction between MC trial and frame is positive and significant, indicating a higher probability of the correct MC trial response in all frames relative to the household bill frame ($\beta_{exoticpets}$ = .296, p = 0.023; $\beta_{vacation}$ = .414, p = 0.001; $\beta_{slotmachine}$ = .267, p = 0.034).

Influence of Priors

Table 2.7 shows how participants in each frame (except the slot machine frame, who were not asked the question) thought real-life costs compared to the monetary sums in the task. The four response options were 'higher in real life', 'lower in real life', 'about the same cost', and 'don't know'. The proportion in the household bills frame who thought the real-life cost was higher (0.15) is substantially lower than the other

	(1)	(2)
	Full Sample	Main Spec.
MC trial (ref. = LC)	-0.510***	-0.522***
	(0.0868)	(0.0876)
Exotic Pet Bills Frame	-0.177	-0.00786
	(0.118)	(0.105)
Vacation Savings Frame	-0.351**	-0.127
	(0.114)	(0.102)
Slot Machine Frame	-0.192	-0.122
	(0.116)	(0.101)
MC*Exotic Pet Bills Frame	0.389**	0.296*
	(0.125)	(0.131)
MC*Vacation Savings Frame	0.492***	0.414**
	(0.121)	(0.128)
MC*Slot Machine Frame	0.309*	0.267*
	(0.123)	(0.126)
0.06 Increment (Medium Difficulty)	-0.213***	-0.255***
	(0.0545)	(0.0580)
0.03 Increment (Most Difficult)	-0.560***	-0.603***
	(0.0617)	(0.0652)
Constant	1.132***	1.189***
	(0.0916)	(0.0830)
Observations	9696	8800

Table 2.6: Mixed Logit Regression table for Experiment 2 Forced-Choice task.

two frames (0.40 and 0.35). The difference is significant at the one-percent level. This may partly explain why the underestimation bias was strongest in the household bill frame. It also points to the vacation manipulation being successful. We set out to create a situation where - if people factored in their experience of vacation costs - it would cause upward adjustment to responses. However, we also intended that any priors for feeding exotic pets would be lower than the sums of the monetary sequences shown, not higher. It is possible people considered the *total* cost of keeping these pets, not just feeding bills as was specified.

	Higher IRL	Lower IRL	Approx. Same	DK
Household Bills	.15	.22	.40	.23
Feeding Exotic Pets	.40	.03	.07	.50
Vacation Savings	.35	.05	.23	.37

Table 2.7: Proportion in different frames responding "higher", "lower", "same", "don't know". Note: IRL above is acronym for "in real life".

Do participants with a prior of "higher in real life" respond differently in the forced-choice task? To test this, we conduct a mixed-effects logistic regression with a binary dependent variable called 'chose more', which codes the trial response as follows: 0 = clicked 'less', 1 = clicked 'more'. The key predictor variable is the prior about real-life costs. Results show that judging the real-life costs to be higher than the sequences shown is a positive and significant predictor of choosing 'more' on a given trial (β = 0.35, p = 0.035). The effect size is about one-quarter the size of MC trial effect, which captures whether clicking 'More' was in fact the correct response (β = 1.46, p < 0.0001). However, these participants were not more discerning about *when* they selected 'more'. In a mixed-effects logistic regression using the MC trials only, the higher-in-real-life prior does not significantly increase the log-odds of recording the correct response (β = 0.24, p = 0.154).

Strategy Adjustment across Task

Strategy adjustment occurred in the forced-choice task, which attenuated the size of the recorded underestimation bias: after approximately 10 trials, there is an uptick in the proportion of 'More' responses (illustrated in Figure 2.12 in Chapter Supplementary Material). Figure 2.9 shows the result of this change: the proportion of MC trials answered correctly rises, but the proportion of LC trials answered correctly falls. This shift in response does not reflect genuine learning, or any kind of improved

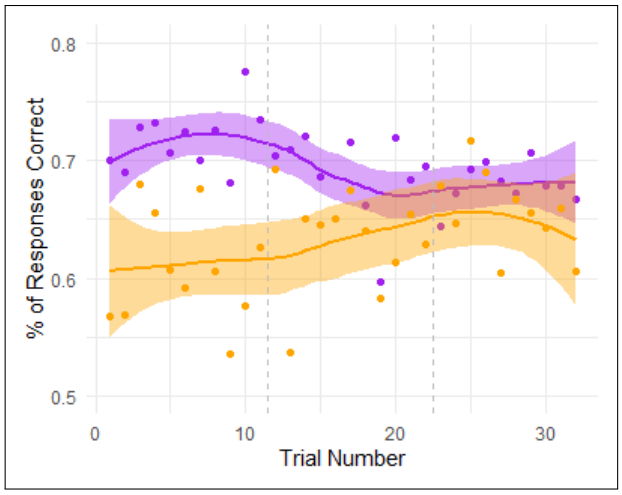

Figure 2.9: Proportion of LC trials (purple dots) and MC trials (orange dots) correct across 32 trials, with matching local regression lines. Underestimation bias is initially larger than the final point estimate suggests.

discernment. Instead, what most likely occurred is that participants realized there would be the same number of LC and MC trials, and adjusted towards the option they had chosen less often. The implication is that taking an average over 32 trials underestimates the bias. Two faint grey dashed lines on Figure 2.9 demarcate the task into thirds. On the first 11 trials, 72% of LC trials are answered correctly, but only 61% of MC trials ($z = 6.26$, Cohen's $d = 0.23$). In the middle third, trials 12-21 inclusive, the gap in performance is 68.5% vs. 63.3% ($z = 2.88$, Cohen's $d = 0.11$). In the final third, the difference is smaller still, 68% vs. 65.5% ($z = 1.54$, Cohen's $d = 0.055$). The effect size halves - then halves again - moving across terciles.

2.3.4 Estimation Strategies

Immediately after completing the task, participants were asked what estimation strategies they used. The typed responses were blind-coded i.e. without knowing how the participant had performed or what frame they were randomly assigned to (though some explanations referenced details of the frame). Four main strategies emerge from analysis of the strategy descriptions, three of which are variations on

rounding. The 'checkpoints' strategy refers to participants keeping track of every time the intuited total passed another salient marker, usually one hundred, but sometimes fifty. Left-digit focus is essentially rounding down. And 'rounding' means rounding up *or* down (direction not specified). In contrast to rounding, just under one-fifth of participants indicated keeping a running average and multiplying.

	Prop.	Mean Error	Error IQR
Average and Multiply	20%	-.049	[-0.13, 0.03]
Checkpoints (Keep Track of 100s)	26%	-.062	[-0.15, 0.02]
Rounding	27%	-.063	[-0.16, 0.02]
Left-Digit Focus	20%	-.056	[-0.13, 0.01]
Other/None	7%	-.071	[-0.16, 0.02]

Table 2.8: Judgment task: Underestimation by strategy type.

	Prop.	P(Correct | LC)	P(Correct | MC)
Average and Multiply	16%	.69	.61
Checkpoints (Keep Track of 100s)	31%	.69	.65
Rounding	23%	.72	.64
Left-Digit Focus	17%	.70	.65
Other/None	13%	.66	.60

Table 2.9: Forced-choice task: Underestimation by strategy type.

The averaging strategy appears to attenuate bias in the judgment task (Table 2.8) but the apparent difference is not significant once other factors are controlled for in regression analysis. Averaging also did not improve forced-choice performance (Table 2.9). Note that 'Prop.' in Tables 2.8 and 2.9 refers to the proportion of the sample who used each strategy.

2.3.5 Simulating Choice Data from Judgment Responses

A related but separate question to whether underestimation bias generalizes is whether it is exacerbated by particular response modes. We simulated choice data

from the judgment responses to allow an approximate comparison of the strength of underestimation by elicitation method. The simulation works as follows: First, each of the six increments is assigned to each of the sequence sums in the judgment task, creating six synthetic referents for each true sum. Then we check whether the typed response is on the 'right side' of each synthetic referent i.e. higher for a negative increment (MC trials), and lower for a positive increment (LC trials). For example, on a trial where the sequence sum was £500 and the increment was -0.09, the synthetic referent would be £455, so a typed response of £460 would be classified as 'correct'. But if the increment was -0.06, the synthetic referent would be £470 and hence the response of £460 would be incorrect.

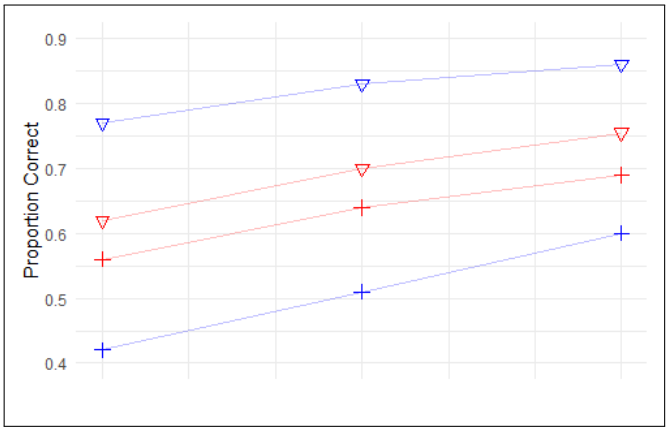

Figure 2.10: Proportion correct for real forced-choice data (red) and choices simulated from judgment responses (blue). '∇' = LC trial, '+' = MC trial. Task difficulty decreasing from left to right.

Results show that the proportion of correct responses on simulated LC trials is 77% for the most difficult trials (increment = 0.03), 83% for medium trials (inc. = 0.06) and and 86% for the least difficult trials (inc. = 0.09). In contrast, on the simulated MC trials, the respective proportions correct are 42%, 51% and 60%. These simulated proportions are shown in blue in Figure 2.10 above. The actual proportions correct from the forced-choice task are plotted in red. As in Experiment 1, the plus symbol represents MC trials, and the downwards triangle the LC trials. The choice data is notably compressed compared to the simulated data. This can partly be explained by choice task respondents building in a 50/50 prior for whether the correct answer

would be 'more' or 'less'. This adjustment may not explain the extent of the disparity however. This is a topic for future research.

2.3.6 Experiment 2 Discussion

In both the judgment and forced-choice tasks, substantial and systematic underestimation bias was recorded. We found no evidence of attenuated underestimation for the household bill frame. In fact, in the forced-choice task, this frame recorded the strongest underestimation. We infer from this that the unhypothesized interactions in Experiment 1 were a consequence of interleaving sequence lengths, or possibly due to random variation in a modest sample size. We do not remark any further on unresolved puzzles in Experiment 1. We focus instead on the general pattern of results.

Experiment 2 tested whether specific contextual factors had a directional effect on estimation error. In drawing inferences about the effect of context, the use of multiple elicitation methods is helpful. Its main benefit is allowing a degree of triangulation. Specifically, in the judgment task, there was a positive correlation between the average estimated annual bill and participant-level mean error. One interpretation is that holding higher priors pushed responses upwards. But the correlation could be spurious - some people might simply have a propensity to type in larger numbers, regardless of the task. However a similar trend was recorded in the forced-choice task: the probability of selecting 'more' on a given trial was significantly higher for those who indicated afterwards that they thought real-life cost of household bills/vacations/feeding exotic pets were *higher* than the sums of the sequences shown on screen. This complements the judgment task evidence and makes a causal interpretation in the first instance more plausible. The subtle effect of priors validated the notion that contextual factors induced by the frame could have an influence on responses.

Lastly, the effect size of underestimation was consistent. The point estimate on mean estimation error was -6.4% in Experiment 1 and -6% in Experiment 2. This is very close to the Scheibehenne (2019) finding of -5.5% mean underestimation. Other close similarities between the studies include the proportion of all sequences that were underestimated: 65% in Scheibehenne (2019) and 67% and 68% in Experiments 1 and 2 respectively. Furthermore, the estimated exponent on the power function was 0.97 in Scheibehenne (2019) and 0.96 in both experiments in this study. These commonalities emerged despite differences in sample demographics and experimental environment.

2.4 General Discussion

We set out to explore the likely impact of underestimation bias in everyday economic contexts, such as estimating the total cost of annual household bills. In such contexts, intuitive summation processes are likely to be used, because dispersion of the pertinent numerical information makes use of analytical processes more cumbersome and time consuming. The results of the multi-experiment study indicate that underestimation bias may affect a wide range of decisions. Three pieces of evidence lead us to this conclusion - namely framing invariance, task invariance, and a consistent effect size. We discuss each point below.

First, framing the sequences in a familiar and meaningful economic context, which might be expected to induce a correct-on-average level of performance, did not eradicate the bias, or even reduce it. A common critique of the external validity of experimental studies is that high stakes and opportunities to learn in the real-world minimize the magnitude of departures from rational decision making (Levitt and List, 2007). Taking this point at face value, we reasoned that if underestimation is partly an artefact of using context-free stimuli, then the use of a familiar, economically meaningful frame might focus attention and induce higher engagement, which could eradicate the bias. But the results supported our primary hypothesis that the

compressive mental number line would be invariant to framing.

Second, both judgment and forced-choice elicitation methods produce the bias. This concordance should not be taken for granted, given the literature on the subtle causes of preference reversal (Cubitt et al., 2004). While all decision tasks have an element of artificiality, forced-choice is arguably closer to real-world decision contexts than a judgment task where one must produce a number. Therefore it was important to test whether underestimation bias would manifest under forced-choice. The broader point is that general psychological phenomena should be detectable regardless of measurement method. While passing this test is not sufficient to attach substantive real-world import to a cognitive bias, it is a necessary condition that reduces external validity concerns. Using both judgment and forced-choice in a complementary fashion is also amenable to field studies, which may be an avenue for future research.

Third, the effect size is consistent with the findings of Scheibehenne (2019), which is supportive of the phenomenon being robust. How one interprets the economic significance of the effect probably depends on the size of the specific monetary stream under consideration. The effect also varies considerably across individuals - one quarter of participants in Experiment 2 underestimated by 10% or more on average in the judgment task. This aspect is relevant to research on protection of vulnerable consumers.

Relatedly, there are several reasons to believe the recorded effect size probably marks a lower bound. For a start, incentivization meant participants were fully attentive. When attention is deprived, numerical mappings tend to be more compressed (Anobile et al., 2012). Underestimation might be exacerbated when households do not need to attend to billing details, for instance when they sign-up to automatic payment, a method which has been shown to increase consumption (Sexton, 2015; Fuentealba et al., 2021), plausibly due to the reduced salience of the bill itself (Gilbert and Zivin, 2014). Second, participants received feedback on their level of absolute accuracy during the judgment task. This may have also reduced the effect size if they

correctly guessed they were undershooting rather than overestimating. This learning would not be possible in a context where a familiar sequence is summed, as objective feedback is unlikely to be available. Even if it is available, learning effects are short-lived without repetition, and familiar summation problems with feedback are unlikely to be conducted frequently. It is hard to envisage how the modest learning in the task would naturally occur in a real-world summation context.

In the forced-choice task, meanwhile, there was no evidence of genuine learning but instead a sophisticated change in behaviour: intuiting that the correct answer would be split evenly between 'less' and 'more' appeared to cause strategic adjustment. This lessened the effect size.[5] But participants did *not* become more precise over the task. Their better accuracy on 'more correct' trials was balanced out by worse performance on 'less correct trials', like a football linesman who realises he has probably missed a few genuine offsides, and over-adjusts to calling offside more often than not - sometimes incorrectly - for the rest of the match.

Turning to applications, at the outset we noted that underestimation could contribute to explaining the enduring puzzle of low switching rates. For instance, consumers may create an approximate internal ordinal ranking of estimated annual expenditures, and consult this ranking when trying to allocate saving efforts efficiently. Underestimation bias will push down the rank of disaggregated expenditures. The sequence length manipulation in Experiment 1 was also motivated by the desire to test decision contexts people regularly encounter in their economic lives. As explained in the introduction, when price information is more spread out, applying an analytical decision process becomes harder. The results were mildly suggestive that underestimation is increasing in sequence length, which is the opposite of the finding in intuitive averaging studies (Tsetsos et al., 2012; Brezis et al., 2015). However future experiments should test this relationship between sequence length and estimation accuracy in a more subtle way. One option would be using an experimental environment that

[5]Future experiments could completely randomize the proportion of correct responses to remove this strategic adjustment by participants, but this might be unnecessarily complex and require participant beliefs about the split to be recorded too.

mimics how people passively observe incoming bills, or perhaps intuitively project the cost of a finance plan for a consumer durable. The preference of digital streaming services for recurring payments is circumstantial evidence that firms are already aware that disaggregating prices may cause the intuited total to be biased downwards. Scholars who speak directly to firms emphasize the point that - unlike rational actors - consumers do not see through pricing plans (Hinterhuber and Liozu, 2017).

The bias also has implications for measurement of economic preferences. Many decision contexts involve streams of monetary amounts, and in such settings underestimation bias confounds straightforward inference of preference parameters from observed choices. This effect may have been hiding in plain sight for some time. Take for example the seminal work of Hausman (1979), which found (inferred) discount rates of over 20 percent in the air conditioner market. It is sensible to consider past usage costs when making an energy investment decision. Underestimating the cost of past usage may contribute to the low willingness to pay for energy efficiency. Regarding more recent research, the fact that underestimation appears to be stronger for longer sequences can help explain why a switch from bi-monthly to monthly billing for water led to an increase in consumption (Wichman, 2017). The bias may inform the ongoing discussion on how best to elicit time and risk preferences, as some methods tacitly assume that intuitive summation is unbiased.

Lastly, underestimation bias may also account for higher-level cognitive biases. For example, it may partly explain the strength of exponential growth bias (EGB) (Stango and Zinman, 2009; Levy and Tasoff, 2017). In order to account for compounding, one must add up the interest in each period. Measurements of EGB have recorded failure to add up the simple interest component or savings contributions fully (Stango and Zinman, 2020; McGowan et al., 2019).

To conclude, this paper has replicated and extended the findings of Scheibehenne (2019), and shown that systematic compression afflicts the process of perceiving and integrating sequential numerical information. Underestimation bias generalizes

across frames and elicitation methods. This bias may affect the outcomes of substantive personal financial decisions and, in some contexts, raise issues of financial capability or consumer protection. The underlying mechanism of a compressive number line is almost certainly adaptive: small numbers were encountered more frequently, and reserving greatest perceptual resolution for the most common stimuli magnitudes is a governing design principle of sensory systems that maximize fitness (Barlow et al., 1961). But this adaptation plausibly causes economic loss in the modern marketplace. A better understanding of when and why the bias arises will help identify choice environments in which mistakes are likely to occur, and also provide evidence on how to present numerical information in a way that minimizes bias in decision-making.

2.5 Supplementary Material

2.5.1 Power Analysis

Experiment 1

The preregistration plan (https://osf.io/rfzgy) did not specifically detail the results of the power analysis that had been performed. It stated:

"One hundred participants will complete the task. Each participant will complete 48 trials (24 in each condition). Collecting 4800 observations provides sufficient power to detect a difference in estimation bias by condition and answer the primary research question (H1)."

Note that there was a mistake here - it should have said *at least* 100 participants would complete the task. Due to unequal attrition, 104 participants were required to have balance across conditions. This wording oversight was corrected in the Experiment 2 preregistration.

Experiment 2

Preregistration text, which is available on OSF at https://osf.io/kc2dy, is as follows:

"We aim to collect data from at least 500 participants. We aim to split the sample 3:2 in favour of the choice task, as binary responses require more observations to meet required power thresholds. We aim for 75 participants in each of the four choice conditions and 50 participants in each of the judgment conditions. This equates to 2400 observations by frame in choice (32 trials per participant) and 1200 by frame in judgment (24 trials per participant). These sample sizes are based on the following power analyses. Note that these estimates are conservative, as the multilevel models used to analyse the data will account for the between-subject component of the

variance more efficiently.

In experiment 1, the effect size for the difference in estimation error by frame for longer sequences was 0.28. Detecting a similar effect size (d = 0.3) with 80% power at the five percent level of significance in a two-tailed t-test with independent means requires 176 observations in each group (GPower software, Faul et al. 2009). To account for the intraclass correlation in the repeated measures, the conservative approach is to multiply the sample size (assuming independence) by the variance inflation factor: VIF = 1 + ρ(m−1), where ρ is the intraclass correlation and m is the size of the cluster. In experiment 1, the intraclass correlation in the judgment task at the id level was estimated as 0.2. The proposed cluster size is 24. This means the inflation factor is VIF = 1 + 0.2 (23) = 4.6. Multiplying the 'independent' group size of 176*4.6 = 809.6. Therefore 1200 observations provides sufficient power to detect an even smaller effect between frames overall.

For longer sequences in the forced-choice task, 72% of 'less correct' trials were answered correctly and 61% of 'more correct' trials. This equates to an effect size of 0.23 (Cohen's d) or an odds ratio of 1.5. The sample size needed to detect this effect with 80% power in a test of two independent proportions, at the five percent significance level in a one-tailed test, is 227 observations per group (GPower 3.1.9.4). To estimate the inflation factor for the forced choice task, we input the intraclass correlation of 0.05 and the cluster size (32). The inflation factor is thus 1 + 0.05(31) = 2.55. This means we need each group size to be approximately 227*2.55 = 681. There are 2400 trials in each condition, equally split between 'more' and 'less' being the correct answer, meaning this test is also sufficiently powered.

Five hundred participants surpasses the minimum sample size requirements to detect an 'underestimation' effect, or, alternatively, is large enough to be confident in a precise null if that is what transpires. The final sample size may vary slightly from 500 due to unequal attrition between conditions, but any deviation will be explained."

2.5.2 Number Sequence Construction

The number sequences were constructed using the Math.random method in Javascript. The following steps outline the processfor Experiment 1. The code will be made available on OSF and GitHub following publication.

1. Six bins of starting sums, spanning £200-800 in total in increments of £10, were hardcoded when the experiment was designed.

2. Each bin was shuffled randomly and two elements from each were selected from each to form the array of 12 starting sums. We will refer to a given element of this list as S-Sum$_i$.

3. For each S-Sum$_i$, six increments were selected randomly from an list of fractions which spanned 0.01 to 0.50 in increments of 0.01. The list was shuffled before each selection. A different set of increments was selected for each S-Sum$_i$. We will refer to a given Fraction as Frac$_j$.

4. For each S-Sum$_i$ and Frac$_j$, $\frac{S-Sum}{12+(12*Frac_j)}$ and $\frac{S-Sum}{12-(12*Frac_j)}$ were calculated, making 12 numbers in total. The rounded figures from these divisions made up the 12-number sequence. The total of the sequence was the target answer for participants.

5. To create the matching six number sequence, the two smallest numbers were added together, and so on, up to the two largest numbers. By design, this procedure meant the pair of length-6 and length-12 sequences which had the same sum also also had identical coefficients of variation (CV). We ran simulations before experiment to ensure this process created sequences with the same range of coefficients-of-variation as in Scheibehenne (2019).

A distinct advantage of this method of constructing sequences is that it minimized the risk that results would be driven by salient numbers, which could happen if all

participants saw the same sequences. Instead, essentially all the sequences shown in the incentivized trials were unique.

2.5.3 Departures from Preregistration Plan

Experiment 1

The preregistration for Experiment 1 was filed on the Open Science Framework (OSF) before data collection began. Link: https://osf.io/rfzgy

The pre-registration plan listed two hypotheses of minor importance that were not discussed in the main text. These were:

H5: Estimation accuracy will decrease as the coefficient of variation of the sequence increases.

H6: We hypothesise that memory-based heuristics - recency, primacy and peak-end effects - will exert a stronger influence on estimates for longer sequences (12 numbers) than shorter ones (6 numbers).

We tested these hypotheses and the results are presented in this Chapter Supplementary Material document.

Another hypothesis, labelled H3 in the original preregistration plan, related to comparing the strength of underestimation across elicitation methods. This analysis was postponed until Experiment 2, when the larger sample size and fully between-subjects design made a more suitable setting for the simulation of forced-choice data from the judgment responses (see main text).

Outliers and Exclusions

Several non-exclusive exclusion criteria (based on accuracy and response time) were proposed but not ultimately implemented. For instance, the skewed distribution of response times made it awkward to use a 2.5 or 3 Standard Deviation accuracy cut off as proposed. The purpose: "to allow for exclusion of mistyped numbers e.g. '5000' or '50' instead of the intended '500"' was achieved using a simpler method of excluding responses that scored less than 50 percent in terms of absolute accuracy. This exclusion criterion was pre-registered and followed for Experiment 2. The main results are robust to trimming the data based on response time by excluding the slowest and fastest five percent of responses or participants.

Experiment 2

The pre-registration for Experiment 2 was filed on the Open Science Framework (OSF) before data collection began. Link: https://osf.io/kc2dy

Analysis

How to simulate choice data from the judgment responses was described in the pre-registration plan: it was originally envisaged that increments would be randomly selected and applied. This random element is the reason why the analysis plan stated that "the simulation will be run 200-1000 times on all valid typed responses." After submitting the preregistration, we realized it was better to apply each of the six increments to *every* judgment response.

Outliers and Exclusions

The simple rule to exclude responses that were less than 50% accurate was followed. An additional criterion, to exclude completely participants who have five or more responses of less than 60% accuracy was also implemented in the analysis shown in the main text (Table 2.4, Column 2). Below we apply robustness tests on the response time dimension by excluding the slowest and fastest responses, as in Experiment 1 (and like in Experiment 1, results do not change).

2.5.4 Experiment 1: Additional Analyses

Effect of Variance on Accuracy

H5 in the pre-registration plan stated that accuracy would be decreasing as the coefficient of variation (CV) increased. The coefficient of variation is the ratio of the standard deviation to the mean. This hypothesis was motivated by experiments that found greater underestimation for higher variance sequences (Olschewski et al., 2021). The mean CV for the monetary sequences in the current study was 0.315 (SD = 0.058). Unreported regression analysis found changes in CV had no effect on estimation error. The simpler Standard Deviation (SD) variable also had a null effect on error when entered as an explanatory variable instead of CV. SD had a significant (negative) effect only when the standardized-sum variable was removed from the specification - recall that underestimation was stronger for larger sequence sums. Removing this standardized-sum variable would be highly misleading however (the correlation between standard deviation and standardized-sum is 0.68).

Primacy and Recency Effects

Estimation of a sequence sum can be strongly influenced by the magnitude of the first or last number(s) that appear because the position of these numbers means they have a disproportionate effect on memory. **H6** hypothesized that these effects would be stronger for longer sequences.

The primacy and recency variables were constructed as follows. First, we calculated the size of each number as a proportion of the sequence sum. To match across different sequences lengths, we added the proportion for the first (last) two numbers in the 12-number monthly sequences to match the proportion for the first (last) number in the six-number bi-monthly sequences. The primacy variable, denoted *PrimacyProp* has the same mean for both sequence lengths ($M_{short} = 0.1671$; $M_{long} = 0.1677$) but

greater dispersion for shorter sequences (SD_{short} = 0.0526; SD_{long} = 0.0365.) The equivalence is present for the *RecencyProp* variable too (M_{short} = 0.1666, SD_{short} = 0.0533; M_{long} = 0.1669; SD_{long} = 0.0356.) Random variation in the two observations from the longer sequences cancels out, which is why SD is lower for longer sequences.

Table 2.10 below shows the output from a regression model. We find evidence of a stronger primacy effect for longer sequences, denoted by the significant coefficient of 0.256 ($p < 0.05$) on the interaction between the 'Longer' dummy and *PrimacyProp* in Column 1. We record a null for recency effects for both short and long sequences (Column 2). In Column 3, we include both primacy and recency variables, and find the significant effect of *PrimacyProp* is robust to inclusion of the recency variable.

	Primacy	Recency	Both
Longer Seq (ref = short bi-monthly seqs.)	-0.0571*	-0.0412	-0.0967**
	(0.0224)	(0.0229)	(0.0347)
PrimacyProp	-0.0312		-0.0331
	(0.0742)		(0.0765)
Longer*PrimacyProp	0.256*		0.285*
	(0.130)		(0.132)
RecencyProp		-0.000762	-0.00972
		(0.0733)	(0.0755)
Longer*RecencyProp		0.161	0.207
		(0.133)	(0.135)
Standardized-Sum	-0.0143***	-0.0143***	-0.0144***
	(0.00271)	(0.00271)	(0.00271)
Bill Frame (ref = abstract)	0.0218	0.0221	0.0222
	(0.0136)	(0.0136)	(0.0136)
Constant	-0.0605***	-0.0657***	-0.0588**
	(0.0158)	(0.0156)	(0.0222)
Observations	2401	2401	2401

Standard errors in parentheses
Stars denote significance as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.10: Effect of sequence presentation on estimation error (Testing H6 in Pre-registration plan).

To aid interpretation of the effect size, we standardized *PrimacyProp* for each sequence length. In the model for longer sequences, the coefficient was 0.0083 (z = 2.10, p =

0.036). In terms of absolute size, a one SD increase in *PrimacyProp* has approximately half the effect on estimation error as a one SD increase in the sequence sum itself (0.0083/0.0145 = 0.57).

Response Time Analysis

Judgment Task

	Full Sample	Main Spec. (T6, Col 3)	RT OBS	RT ID
Bill frame	0.0115	0.00578	0.0137	0.0142
	(0.0159)	(0.0144)	(0.0161)	(0.0168)
Longer	-0.0408 ***	-0.0291 ***	-0.0456 ***	-0.0396 ***
	(0.00900)	(0.00766)	(0.00950)	(0.00937)
Billframe*Longer	0.0388 **	0.0302 **	0.0436 ***	0.0413 **
	(0.0127)	(0.0109)	(0.0136)	(0.0135)
Standardized Sum	-0.0214 ***	-0.0142 ***	-0.0211 ***	-0.0210 ***
	(0.00314)	(0.00270)	(0.00335)	(0.00334)
Choice Task First	0.0239	0.0264 *	0.0279	0.0302 *
	(0.0146)	(0.0134)	(0.0147)	(0.0154)
Constant	-0.0695 ***	-0.0716 ***	-0.0729 ***	-0.0760 ***
	(0.0134)	(0.0121)	(0.0136)	(0.0142)
Observations	2496	2401	2246	2256

Table 2.11: Response time exclusions in Experiment 1 judgment task.

The mean response time in the judgment task was 9.6 seconds (SD = 10.04) and the median was 6.2 seconds. At the participant level, mean response times on incentivized trials varied from 3.1 to 41 seconds, with the interquartile range bunched at 5.6-10.9 seconds. Table 2.11 above shows the robustness of the main results to different response time exclusion criteria. To aid comparison, the first column shows the full sample, and the second column shows the main specification from the text in Table 2.1. Column 3 includes the only the middle 90 percent of observations by response time, i.e. the slowest and fastest five per cent have been trimmed. Compared to Column 2, the effect size for three variables are slightly stronger (the dummy for longer sequences, the Standarized Sum variable, the dummy for task order), but the

overall pattern is unchanged. The same holds for Column 4, where the slowest and fastest five per cent of respondents (judged by mean response time) are excluded. Overall the results emphasized in the main text are robust to alternative exclusion criteria based on response time.

The different trial types were not associated with differences in response time, with the exception of the sequence length manipulation, with response times on longer trials 600ms longer (p = 0.043). Response time decreased as the task progressed.

Forced-Choice Task

	(1)	(2)	(3)	(4)
	Full	Restricted	90% Obs. by RT	90% IDs by RT
MC Trial (ref. = LC)	-0.141	-0.140	-0.179	-0.192
	(0.155)	(0.166)	(0.163)	(0.161)
Bill frame	-0.340*	-0.337*	-0.358*	-0.339*
	(0.153)	(0.139)	(0.160)	(0.156)
MC trial*Bill Frame	0.415*	0.393*	0.389*	0.400*
	(0.177)	(0.191)	(0.187)	(0.184)
0.06 Increment (ref. = 0.09)	-0.218	-0.206	-0.253*	-0.199
	(0.112)	(0.122)	(0.119)	(0.116)
0.03 Increment	-0.747***	-0.849***	-0.844***	-0.741***
	(0.109)	(0.117)	(0.116)	(0.113)
Longer	0.209	0.196	0.246	0.199
	(0.128)	(0.138)	(0.136)	(0.134)
MC*Longer	-0.545**	-0.520**	-0.571**	-0.531**
	(0.179)	(0.192)	(0.190)	(0.187)
Trial Number	0.0156*	0.0157*	0.0173*	0.0169*
	(0.00638)	(0.00688)	(0.00678)	(0.00667)
Forced-Choice Completed First	0.200	0.256**	0.189	0.180
	(0.122)	(0.0985)	(0.127)	(0.123)
Constant	0.959***	1.075***	1.014***	0.933***
	(0.174)	(0.172)	(0.182)	(0.178)
Observations	2496	2184	2246	2256

Table 2.12: Experiment 1 Forced-Choice Results with Response Time Exclusions

The mean response time in the forced-choice task was 4.3 seconds (SD = 6.1) and the

median was 2.7 seconds. The interquartile range was 1.98 to 4.37 seconds. Table 2.12 above shows the sensitivity of the forced-choice main results to different response time exclusion criteria. The format is the same as Table 2.11, with Columns 1-4 respectively showing the full sample, the preferred specification in the main text, the results when the model is run only on the middle 90% of observations based on response time, and finally the middle 90% of participants based on mean response time. The results are not sensitive to these response time restrictions.

	(1)	(2)	(3)	(4)
	Full	Restricted	90% Obs. by RT	90% IDs by RT
MC trial	-0.203 *	-0.206 *	-0.269 **	-0.260 **
	(0.0881)	(0.0951)	(0.0935)	(0.0920)
0.06 Increment (ref = 0.09)	-0.226 *	-0.213	-0.260 *	-0.205
	(0.111)	(0.122)	(0.118)	(0.116)
0.03 Increment	-0.757 ***	-0.857 ***	-0.854 ***	-0.751 ***
	(0.108)	(0.117)	(0.115)	(0.113)
Bill frame	-0.129	-0.140	-0.157	-0.131
	(0.122)	(0.0975)	(0.127)	(0.123)
Longer	-0.0709	-0.0682	-0.0508	-0.0783
	(0.0881)	(0.0950)	(0.0935)	(0.0919)
Forced Choice Completed First	0.205	0.260 **	0.192	0.183
	(0.122)	(0.0980)	(0.127)	(0.123)
Trial Number	0.0160 *	0.0158 *	0.0178 **	0.0174 **
	(0.00636)	(0.00687)	(0.00676)	(0.00665)
Constant	0.979 ***	1.102 ***	1.048 ***	0.955 ***
	(0.161)	(0.156)	(0.170)	(0.165)
Observations	2496	2184	2246	2256

Table 2.13: Forced-Choice task results without interaction terms.

Table 2.13 above shows the results for the regression specification without interaction terms. Columns 3 and 4 show underestimation bias - which implies a lower probability of a correct response on MC trials - is stronger under response time exclusions. Otherwise the pattern of results is unchanged; the main finding is not driven by a subset of respondents who responded particularly quickly or slowly.

No trial types had a significantly different response time. The estimate on the 'Longer' dummy variable was positive and approximately 400 ms, but fell shy of significance

(p = 0.053). Adding trial number showed that as the task progressed, participants responded faster, but the effect size was modest with an estimate of approximately 60ms per trial ($\beta_{trialnumber}$ = -63, p < 0.0001).

Forced-Choice Task: Additional Analyses

LPM Results

	Main Spec. LPM
MC Trial	-0.0254
	(0.0330)
Bill frame	-0.0657*
	(0.0276)
MC trial*Bill frame	0.0772*
	(0.0385)
0.06 Increment (ref. = 0.09)	-0.0379
	(0.0236)
0.03 Increment	-0.178***
	(0.0236)
Longer	0.0391
	(0.0273)
MC*Longer	-0.108**
	(0.0386)
Trial Number	0.00318*
	(0.00139)
Forced-Choice Completed First	0.0520**
	(0.0198)
Constant	0.737***
	(0.0340)
Observations	2184

Table 2.14: Mixed-effects Linear Probability Model for Experiment 1 Forced-Choice Task.

Table 2.14 above shows the results of applying a Linear Probability Model to the forced-choice task data. The pattern of results is the same as Logistic Model. Coefficients denote percentage point changes in probability of recording correct response.

Frame-Specific Order Effect

Table 2.15 shows the percentage correct (rounded to one decimal place) for each combination of frame, trial type (LC and MC) and task order. The difference in the first column stands out and is highly significant (two-sample test of proportions, $z = -2.59$, $p < 0.01$). It is also significant in a regression model that includes a three-way interaction between the task order, trial type and frame dummy variables.

	Familiar Bill Frame		Abstract Frame	
	LC	MC	LC	MC
Forced Choice Task First	70.7	66	73.1	66
Judgment Task First	60.8	64.8	71.8	61.9
PP Gap	9.9	1.2	1.3	4.1

Table 2.15: How task order had differential effect on accuracy depending on frame and trial type.

One potential cause of this isolated order effect is that participants continued to apply a strategy they developed in the judgment task, but in the choice task it was disadvantageous. We are disinclined to speculate given the small samples behind these sub-analyses. However, without noting this order effect, the difference between the frames might be overstated.

Strategy Adjustment Across Task

Figure 2.11 below shows how the proportion of 'More' responses changed across the task: it increased at the task halfway point (12 out of 24 trials). This may be evidence of strategy adjustment on the part of the incentivized participants, on realizing that the correct answer might well be split 50:50 between 'Less' and 'More' (it was). The same pattern of strategy adjustment occurred in Experiment 2 (see Figure 2.12)

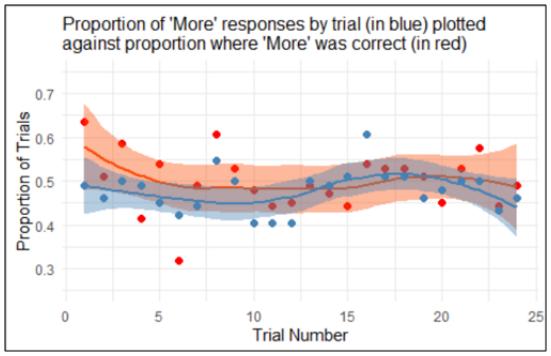

Figure 2.11: The red line is the proportion of trials where 'more' is the correct answer. It averages exactly 50%. The blue dots indicate the proportion of 'more' responses by trial. There is an uptick around the halfway mark (trial 12).

2.5.5 Experiment 2: Additional Analyses

Primacy and Recency Effects

	(1)	(2)	(3)
	Primacy	Recency	Both
Exotic Pet Bills Frame	0.0157	0.0160	0.0157
	(0.0165)	(0.0165)	(0.0165)
Vacation Savings Frame	0.00423	0.00460	0.00425
	(0.0162)	(0.0162)	(0.0162)
Slot Machine Frame	0.00306	0.00322	0.00312
	(0.0165)	(0.0165)	(0.0165)
Standardized Sum	-0.0133 ***	-0.0133 ***	-0.0133 ***
	(0.00179)	(0.00180)	(0.00179)
PrimacyProp	0.226 **		0.222 **
	(0.0698)		(0.0701)
RecencyProp		-0.0617	-0.0423
		(0.0686)	(0.0688)
Constant	-0.0834 ***	-0.0596 ***	-0.0796 ***
	(0.0132)	(0.0132)	(0.0146)
Observations	4560	4560	4560

Table 2.16: Primacy and recency effects in Experiment 2 Judgment Task

As in Experiment 1, a significant primacy effect was recorded. No recency effect was detected. Table 2.16 above shows the primacy effect in Column 1, the null recency

effect in Column 2, and both variables are included in Column 3.

Response Time Analysis

Judgment Task

Table 2.17 shows the robustness of the main results to exclusions based on response times. The first column shows the preferred specification from the main text for comparative purposes. Column 2 drops the slowest and fastest five percent of responses; Column 3 does this at the ID level. The change in coefficients is relatively smaller than in Experiment 1, probably due to the larger sample size.

	Main Spec.	RT-middle 90% of obs.	RT -Middle 90% of IDs
Exotic Pet Feeding Bills	0.0161	0.0186	0.00822
	(0.0165)	(0.0168)	(0.0169)
Vacation Savings	0.00459	0.00353	-0.00558
	(0.0162)	(0.0165)	(0.0170)
Slot Machine	0.00314	0.00267	-0.000354
	(0.0165)	(0.0169)	(0.0169)
Standardized Sum	-0.0132 ***	-0.0140 ***	-0.0139 ***
	(0.00180)	(0.00192)	(0.00190)
Constant	-0.0647 ***	-0.0661 ***	-0.0632 ***
	(0.0119)	(0.0121)	(0.0122)
Observations	4560	4101	4121

Table 2.17: Experiment 2 Judgment Task: Sample restricted based on response time to middle 90 percent of observations (Column 2) and participants (3). Results do not change.

Unreported analysis investigated trial features associated with differences in response time. There was no difference in response time across frames. Interestingly, a one standard deviation increase in the sequence sum was associated with an increase in response time of about 400ms ($z = 3.57$, $p < 0.0001$). There was also a strong negative relationship between trial number and response time, but the effect size was minute ($\beta_{trialnumber} = -103$, $z = -6.12$, $p < 0.0001$).

Forced-Choice Task

Participants' mean response time was 3.8 seconds (SD = 6.44) and the median response time was 2.45 seconds. The inter-quartile range was 1.83-3.70 seconds. Table 2.18 shows the robustness of the main result to trimming the sample based on response time, as in Experiment 1. The specification from the main text is reproduced in Column 1 for comparative purposes. Column 2 reports results for when the upper and lower five per cent of responses have been trimmed. Column 3 reports results for the sample trimmed at the ID-level. The main effect of interest is the negative coefficient on the MC dummy variable, which signals an underestimation bias. This coefficient is significant at the one percent level in all specifications.

	Main Spec.	RT - middle 90% of obs	RT-middle 90% of IDs
MC	-0.522 ***	-0.652 ***	-0.500 ***
	(0.0876)	(0.0928)	(0.0939)
MC*Exotic Pet Bills	0.296 *	0.364 **	0.366 **
	(0.131)	(0.139)	(0.140)
MC*Vacation Savings	0.414 **	0.358 **	0.408 **
	(0.128)	(0.136)	(0.136)
MC*Slot Machine	0.267 *	0.334 *	0.277 *
	(0.126)	(0.133)	(0.133)
0.06 Increment (ref. = 0.09)	-0.255 ***	-0.229 ***	-0.270 ***
	(0.0580)	(0.0613)	(0.0616)
0.03 Increment	-0.603 ***	-0.656 ***	-0.660 ***
	(0.0652)	(0.0687)	(0.0690)
Constant	1.189 ***	1.242 ***	1.187 ***
	(0.0830)	(0.0885)	(0.0876)
Observations	8800	7939	7840

Table 2.18: Experiment 2 Forced-Choice Task: Excluding observations based on response time. The base effect of each frame has not been included due to space constraints.

Forced Choice Task: Additional Analyses

LPM Results

	Main Spec. LPM
MC trial	-0.115***
	(0.0190)
Exotic Pet Bills Frame	-0.00198
	(0.0220)
Vacation Savings Frame	-0.0264
	(0.0217)
Slot Machine Frame	-0.0253
	(0.0214)
MC*Exotic Pet Bills Frame	0.0674*
	(0.0281)
MC*Vacation Savings Frame	0.0917**
	(0.0276)
MC*Slot Machine Frame	0.0587*
	(0.0273)
0.06 Increment (ref = 0.09)	-0.0527***
	(0.0121)
0.03 Increment	-0.132***
	(0.0140)
Constant	0.768***
	(0.0172)
Observations	8800

Table 2.19: Mixed-effects Linear Probability Model for Experiment 2 Forced-Choice Task. The pattern of results is the same as Logistic Model. Coefficients denote percentage point changes in probability of recording correct response.

Strategy Adjustment Across the Task

Figure 2.12 below shows the proportion of 'More' responses increased just before the task halfway point. The pattern is almost identical to Figure 2.11 which illustrates the same trend for Experiment 1. One result of the increased tendency to respond 'More' was lower precision on LC trials, where the correct answer was 'Less', as shown by Figure 2.9 in the main text. Without this adjustment, the estimated underestimation bias would likely have been larger.

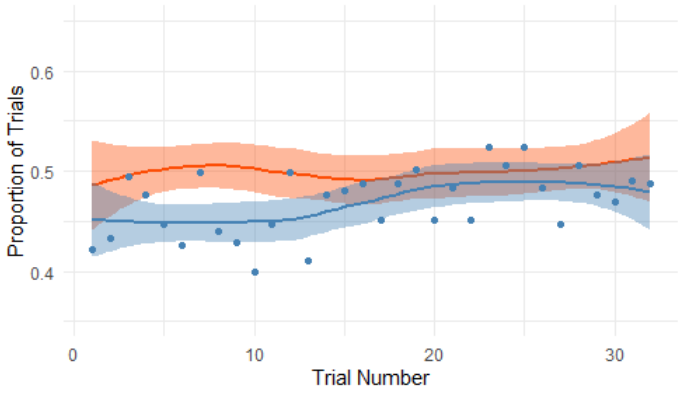

Figure 2.12: The red line is the proportion of trials where 'more' is the correct answer. It averages exactly 50%. The blue dots indicate the proportion of 'more' responses by trial. There is an uptick after 10 trials and before the halfway point (trial 16).

2.5.6 Experimental Instructions

Figure 2.13: Introductory Screens for Familiar Household Bill frame: Left panel – short introduction and picture of logos (first screen). Right panel - general instructions for two tasks to follow (second screen).

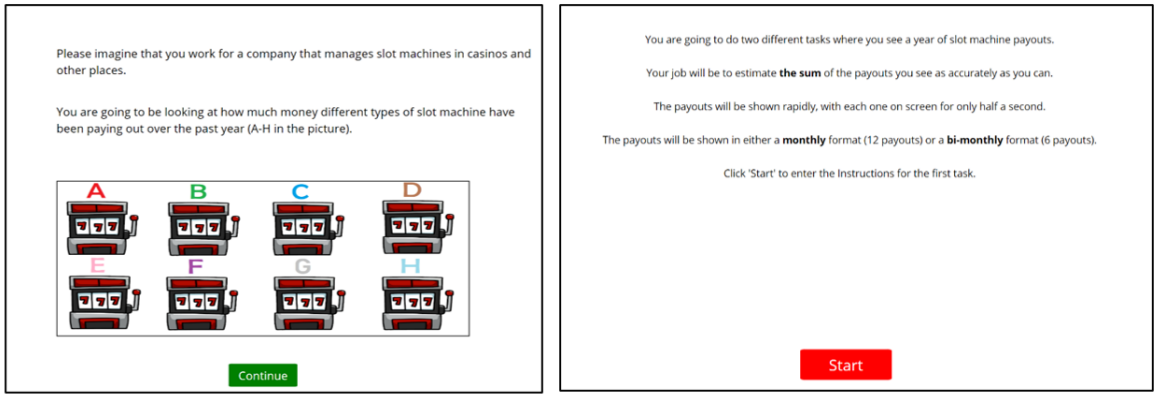

Figure 2.14: Introductory Screens for Abstract slot machine frame: Left panel – short introduction and picture of cartoon slot machines (first screen). Right panel - general instructions for two tasks to follow (second screen).

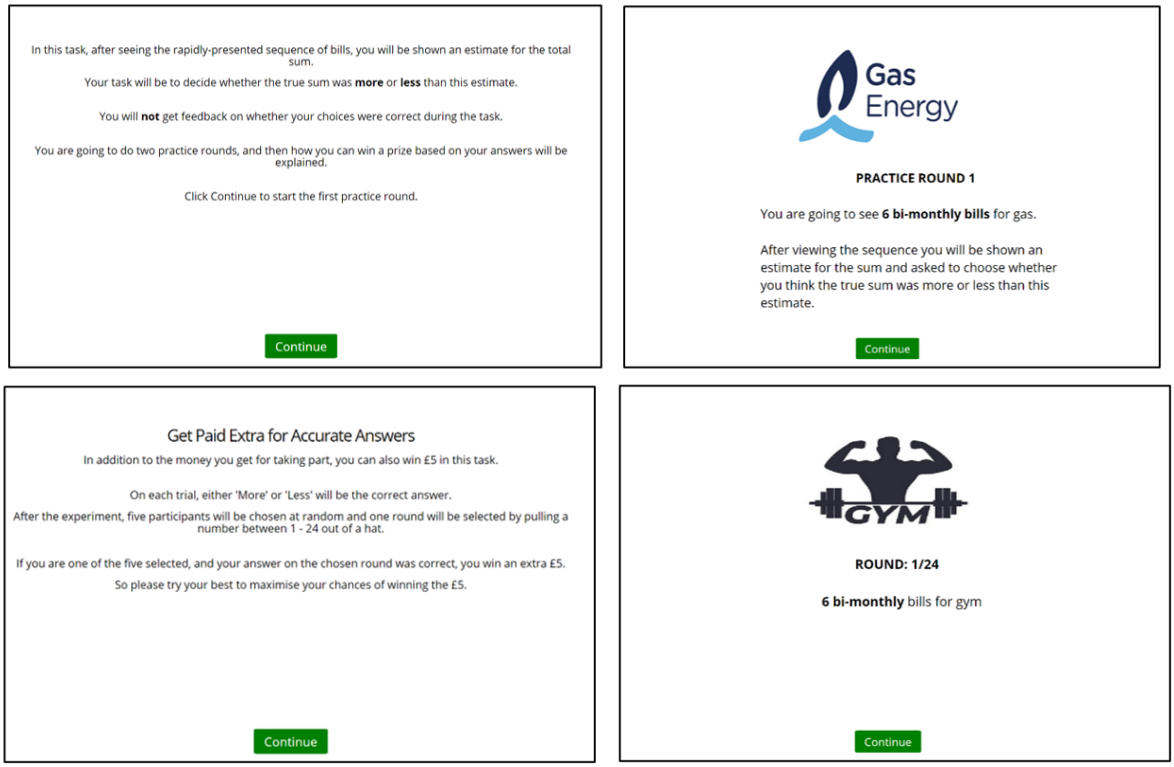

Figure 2.15: Forced-Choice Task Instructions. Top left panel: Introductory text (screen 1). Top right panel: Practice Trial 1 (screen 2). Bottom left panel: Explanation of random lottery incentive. Bottom right panel: Example of what first incentivised trial looked like - Note that this was varied at the ID level (it was equally probable to get a bi-monthly or monthly bill for any of the eight services). Also note that a screenshot of the response screen is shown in the main manuscript (Figure 2.1).

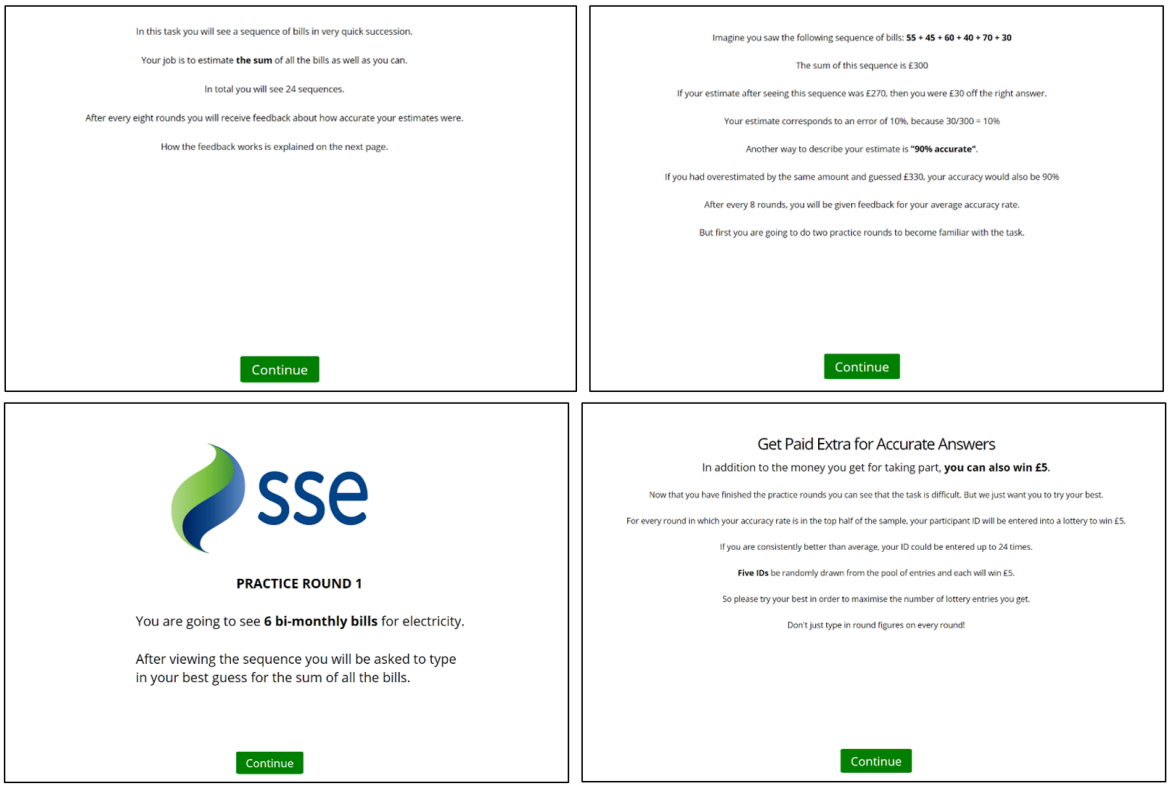

Figure 2.16: Judgment Task Instructions. Top left panel: Introductory text (screen 1). Top right panel: Explanation of feedback on absolute accuracy level (screen 2). Bottom left panel: Practice round 1. Bottom right panel: Explanation of performance-weighted lottery incentive. Note that the real trial introduction screen is not shown as it was identical to forced-choice introduction screen above.

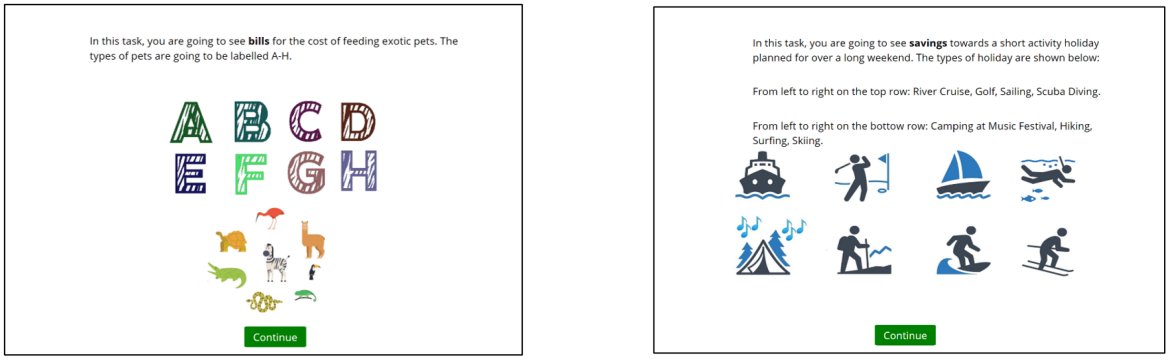

Figure 2.17: Two additional conditions in Experiment 2. Left panel: Introductory Screens for Bills for Feeding Exotic Pets Frame. Right panel: Introductory Screens for Holidays Frame

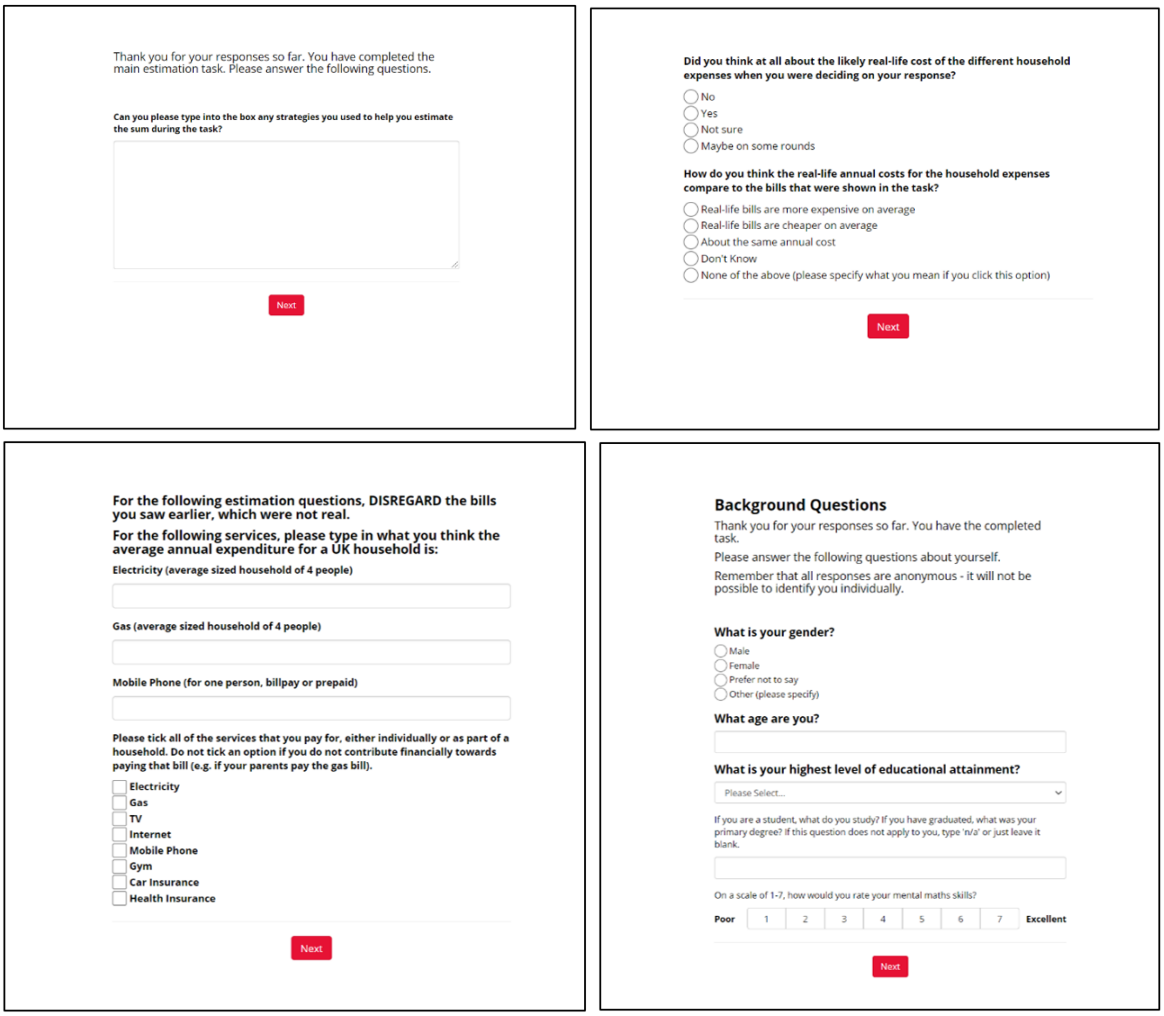

Figure 2.18: Post-Task Questionnaire: Top left panel: Strategy open text response. All participants were shown this. Top right panel: Questions to elicit priors, and whether they were used in the task. Not shown to abstract slot machine frame participants. Bottom left panel: Numeric priors for some bills – shown to household bill frame only. Bottom right panel: Demographics questions – shown to all participants

Chapter 3 Looking beyond time preference: testing cognitive mechanisms relevant to the energy paradox

Abstract

Time preferences are considered a leading cause of the energy efficiency gap. We test two novel mechanisms that are distinct from time preferences, but which produce identical behaviour in many common investment settings, where costs are paid upfront and benefits are spread over time. In a pre-registered experiment on a large, nationally representative sample of car buyers, we elicited willingness-to-pay (WTP) for an improvement in fuel economy between two cars which were otherwise identical. The payment schedule for the investment was varied, as was the temporal framing of fuel costs for both cars. How this information was presented on the label was varied between-subjects. Results supported the preregistered hypotheses, with annualised WTP increasing as the payment schedule for the investment was spread out over time, and WTP lower when the monetary saving on fuel was shown over shorter periods of time (e.g. monthly instead of annual). Findings indicate that the proportion of the energy efficiency gap attributed to present-biased preferences may be exaggerated.

This chapter is co-authored with Prof. Eleanor Denny and Prof. Pete Lunn, who collaborated on the design and gave helpful feedback on the analysis and write-up. These findings were accepted to SJDM 2021 but not presented due to conference postponement, and have been accepted to EAERE 2022 for an oral presentation.

3.1 Introduction

Adoption of greener technologies is crucial to avert a climate crisis (IEA, 2018). Transportation is a key sector for such adoption because it is one of the largest contributors to carbon dioxide emissions worldwide (IPCC, 2021). But incentivizing adoption through Pigouvian taxes, the method preferred by economists, is politically fraught. The reluctance to make polluters pay for carbon means 'nudging' consumers towards adoption is essential to bridge the energy-efficiency gap (Gerarden et al., 2017). However, informational nudges must be designed with precise knowledge of *why* existing behaviour occurs in order to be effective. At the same time, there is no consensus on the primary determinants of the energy-efficiency gap. Leading explanations include imperfect information, inattention, bounded-rationality, and short-sightedness or 'myopia'- in other words a high subjective discount rate (Gerarden et al., 2017). This last factor, a high discount rate, was the original explanation for the energy paradox (Hausman, 1979). This paper investigates the explanatory power of two novel mechanisms which are completely distinct from the predominant time preference explanation (and also distinct from simple inattention). Crucially, both mechanisms can produce behaviour that appears identical to present-biased choices.

The first mechanism is 'concentration bias', which is a feature of the focusing model of economic choice (Kőszegi and Szeidl, 2013). Concentration bias describes how in a multi-attribute choice setting, greater utility weight is given to attributes with larger differences, because larger differences attract attention. According to concentration bias, incremental savings on fuel costs are underweighted because the magnitude of fuel cost differences is much smaller than the magnitude of the upfront price difference. People focus their attention on the larger difference and this distorts the expected utility calculation. However, there is no misperception of the actual costs of each option over time. Allcott (2016) suggested testing concentration bias as a potential cause of the energy paradox in his review of the literature on energy

efficiency.[1] This paper, to the best of our knowledge, is the first to implement this suggestion.

The second alternative to time preference that we test is a systematic tendency to underestimate the sum of numerical sequences when people approximate the total intuitively. In other words, when people approximate a total in their heads, either by multiplication or addition, that estimated total tends to be less than the true sum on average. Intuitive summation generating an *underestimation bias* is a recent finding in cognitive psychology (Scheibehenne, 2019). A compressive internal scaling of numerical quantities is the most likely cause of this bias (Dehaene et al., 2008), which is only beginning to receive attention from economists. Recent work indicated that underestimation bias becomes stronger as the sequence to be summed increases in length. This suggests that underestimation bias might be more pronounced for consumer durables like cars, where savings from investing in greener technology accumulate over a long period of time.

Discounting and underestimation bias can cause the same economic behaviour. The following example illustrates the point: imagine a car buyer is choosing between two cars which are identical in every way, except one is more fuel-efficient. The buyer's expectations for annual mileage and fuel prices are well calibrated. Under these beliefs, assume the more fuel-efficient car will save her €42 per month. She expects to own the car for exactly three years. What is her additional WTP for the better car? Suppose her annual discount rate is five per cent (so $\delta = 0.95$) and for simplicity she evaluates savings on the final day of each year, then her present discounted value is: $\delta(42*12) + \delta^2(42*12) + \delta^3(42*12) =$€1365. Now assume instead that $\delta = 1$ i.e. she cares equally about the future and present. However, she exhibits underestimation bias, with intuited savings underestimated by 10%. She sets her WTP valuation equal to this perceived sum: (€42*36)*0.9 = €1360.

[1] "Could bias toward concentration be relevant for some energy efficiency decisions? Because the benefits of energy efficiency come in small flows every time an electricity bill is paid or the gas tank is filled, this is a natural application of Koszegi & Szeidl's (2013) model. Even a lab experiment framed in an energy use context would be interesting if it measured and distinguished bias toward concentration from other potential mistakes." (Allcott, 2016, p. 27).

We employed a willingness-to-pay (WTP) experiment to test the explanatory power of concentration bias and underestimation bias. The experiment sample was large and nationally representative (n = 2368), and all participants had bought a car in the past five years. The hypothetical scenario was very similar to the earlier WTP example. Participants saw two cars, which were identical except for fuel efficiency, and expressed how much extra they would pay for the more fuel efficient car (hereafter 'greener' car for short). The setting was leasing a plug-in hybrid vehicle for three years. The lease contract took the form of a upfront deposit and 36 monthly payments.

Three factors matter for determining the WTP of an energy-efficiency investment. The first two are its monetary benefit and cost. A rational agent will apply time preferences over the flow of benefits and costs, but the underlying components of total cost and total benefit will be perceived correctly. In contrast, how a boundedly-rational consumer perceives the total benefit or cost will depend on the nature of the extrapolation from the stated per period benefit/cost. A distortion in this mapping process may reinforce or counteract time preferences. Both concentration bias and underestimation bias act through this channel. We manipulate the dispersion of costs and benefits in different treatments so that concentration and underestimation bias predictions can be separated from those of time preference. The third factor that determines WTP relates to how information is presented: these 'choice architecture' manipulations engage context-dependent preferences, often by increasing the relative salience of certain attributes, which affords them greater decision weight (Bordalo et al., 2012), or by making more accessible information which would otherwise require cognitive effort to draw out (Kahneman, 2011).

These three dimensions were varied between-subjects. On the benefit side dimension, we varied the temporal frame of the expected fuel savings from the greener car (per month, per year, lifetime (three-year)). We hypothesised that longer temporal frames of savings would increase WTP. Time preferences alone predict no difference in WTP across framings because the framing does not alter *when* the saving is realized.

On the costs side, we varied the structure of the additional payment for the greener car. One third of subjects were asked: "How much extra would you be willing to pay per month?" The two other groups were asked to give their additional annual payment, or additional once-off payment as part of the deposit, respectively. The predominant time preference functional forms, namely exponential and hyperbolic discounting, predict that WTP should be increasing as a greater proportion of the cost is deferred to the future. This means the additional once-off deposit payment treatment should have the lowest total WTP. Exponential and β-δ behavioural models of time preference make no clean prediction over whether WTP should be higher in the annual or monthly payment treatment, because the scenario did not specify when in the year the additional annual payment took place. In contrast, both concentration bias and underestimation bias plainly predict that the monthly cost treatment should have higher WTP. The choice architecture manipulations altered the accessibility of the monetary saving from the greener car and varied how the saving was presented within the personalised fuel forecast. (The motivation for these manipulations is given in Section 3.)

We also designed a separating test to measure the relative predictive power of concentration bias and underestimation bias. The details of this test are described in Section 3, but the logic is as follows: We calculated the pattern of WTP responses between different combinations of payment schedule and fuel cost frame that each of the two biases would predict. We pre-registered that whichever pattern emerged, relatively greater predictive power would be attributed to this bias. This is an indirect approach, which was made necessary by the short time period available to conduct the experiment. In comparison, a direct approach would first obtain independent measures of each bias at the individual level, and use these measures to predict WTP responses. At first glance, the results of the indirect separating test appeared to favour concentration bias. However, closer inspection showed that a prerequisite condition shared by both patterns was not met, making the separating test inconclusive.

Results showed that framing the fuel cost over a longer period of time significantly

increased willingness-to-pay, with a median point estimate of 13% for the annual and three-year fuel cost frames relative to the monthly frame. This suggests that when people only see the monthly cost, they underestimate accumulated fuel savings. WTP was also significantly higher when the additional payment was more disaggregated, with an estimate of WTP being 35% higher in the monthly payment condition.

The rest of the paper proceeds as follows. Section 2 reviews the most relevant literature and presents a simple decision framework for how the outlined mechanisms affect purchase decisions. Section 3 describes the experiment design, hypotheses and procedure. Section 4 presents the results. Section 5 discusses the findings and the implications for theory and policy, and proposes directions for future research.

3.2 Related Literature and Decision Framework

This section first provides background context on the range of factors that plausibly contribute to the energy-efficiency gap, then summarises the most relevant literature on (i) the mechanisms we test and (ii) the effect of labels on valuation of energy-efficiency in a range of consumer durables.

3.2.1 Leading Causes of the Energy-Efficiency Gap

Numerous causes have been forwarded as contributing to the energy-efficiency gap. Apart from the aforementioned time preferences and cognitive biases, other factors are: limited information, financial constraints, attention deficits, preferences for other appliance attributes, principal-agent problems[2], uncertainty over future energy prices (and hence likely savings), and regulations that distort the prices people face. These causes have been categorised into market failures and "behavioural failures"

[2]And other instances of misaligned incentives which are not classic principal-agent problems, for example when a landlord has no reason to improve the insulation of a property because the heating bill is paid by the tenant.

(Gillingham and Palmer, 2020; Allcott, 2016). Reviewing potential market failures is beyond the present scope but Gerarden et al. (2017) provide an excellent treatment of the issues.

The leading framework for categorising departures from the rational agent model is the DellaVigna (2009) taxonomy of non-standard preferences, non-standard beliefs and non-standard decision-making. The latter category includes all quirks, mistakes, shortcuts, misconceptions and other decision-making idiosyncrasies that an optimising rational agent would not succumb to, but may impact outcomes in markets populated by boundedly-rational consumers. Two features of this category, namely inattention and salience effects, are the most important for the energy-efficiency gap according to Gerarden et al. (2017). For instance, Turrentine and Kurani (2007) interviewed recent car buyers and found that many were inattentive to fuel costs when buying a car, and others considered it in a very simple way that did not reflect calculation of net-present value. The labelling interventions reviewed in the next section essentially attempt to counteract the limited attention this decision attribute receives.

Regarding non-standard preferences, time-inconsistent preferences that arise from a quasi-hyperbolic discount function (Laibson, 1997) are considered relevant to the energy-efficiency gap, because they capture the self-control problem between wanting to reduce energy usage in the future but being unwilling to incur some sacrifice in the present (Bradford et al., 2017). Another non-standard preference is loss aversion, which has been posited as a reason why positive net-present-value investments might be willingly foregone: If consumers are loss averse, and future energy prices are uncertain, they may weigh the potential negative payoffs heavily enough to be deterred from purchase, even though they know the investment would probably have positive net benefits (Gillingham and Palmer, 2020). However, although theoretically plausible, the empirical evidence for loss aversion as a key wedge in the energy efficiency gap is weak (Gerarden et al., 2017).

There is evidence that beliefs over energy usage may be systematically misaligned. Attari et al. (2010) showed that consumers display a 'central tendency' in their perceptions of energy use, with severe underestimation of the energy usage from intensive sources such as heating and cooling, and minor overestimation for low-intensity products like a light bulb. Larrick and Soll (2008) documented the concept of the miles-per-gallon (MPG) illusion: people's intuition is that fuel costs scale linearly in MPG, which is not the case, as the relationship is strongly non-linear. Of course, the EU standard of "litres per 100km" provides consumers with a linear relationship, though people may still underestimate fuel cost differences between vehicles if they underestimate their mileage.

3.2.2 Experimental Evidence on Labelling and Energy Valuation

Undervaluation of energy efficiency is inferred from heightened responsiveness by consumers to a change in upfront costs relative to a change in future costs. This is often referred to as myopia, a term arguably too broad to be useful.[3] There is a lack of consensus on the degree to which consumers are myopic (Greene, 2010). The issue remains contested, as illustrated by two recent labelling experiments. A natural experiment leveraged a change in fuel labels, after the fuel economy of two popular small cars had been overestimated and subsequently corrected (Gillingham et al., 2021). The change in demand implied that consumers were indifferent between $1 in discounted fuel costs and 15-38 cents in the vehicle purchase price (assuming discounting at 4%), which is stronger undervaluation than typically found by exploiting changes in fuel prices (Allcott and Wozny, 2014). Altering the assumptions about projected annual mileage, length of ownership, expectations over future fuel costs, and owners' discount rates changed the finding, but under most plausible combinations, undervaluation was substantial.

[3]Myopia covers a spectrum of departures from the behaviour of a rational agent - imperfect information, inattention, deliberation costs, high discount rates, and so on. Silvi and Rosa (2021) provide a comprehensive guide.

In contrast, no undervaluation was recorded in a large-scale labelling intervention experiment (in both the field and online) that drew buyers' attention to fuel economy (Allcott and Knittel, 2019). The paper concluded that imperfect information and inattention do not have a significant systematic effect on vehicle markets after finding precise null effects for treatments. The implication is that fuel economy receives a low decision weight rather than being undervalued. However, Allcott and Knittel (2019) entertained the possibility that the labels were not effective in *actually* making consumers factor in fuel costs to their decision. On this point, it is worth noting that the labels included information on what else the lifetime monetary saving could purchase in terms of clothes or airline tickets.[4] This aspect of the label may have distracted participants from comparing the saving against the upfront price difference, or calculating the saving under a shorter ownership period and comparing that total to the price difference. This null finding in a high-stakes decision stands in contrast to the wider literature on energy efficiency labelling interventions for consumer durables. Generally, WTP is higher for energy-efficient appliances when savings are framed in monetary terms, rather than in terms of energy output (Newell and Siikamäki, 2014; Min et al., 2014; Andor et al., 2020). WTP is also often higher when lifetime energy costs are shown on the label (Heinzle, 2012) but the field evidence suggests the success of this intervention may depend on the complementarity of staff training quality (Kallbekken et al., 2013; Carroll et al., 2016).

When it is possible that valuations are driven by cognitive limitations, rather than intrinsic economic preferences, it is worthwhile to design tasks that allow no room for preferences. This approach was taken in an online experiment that presented participants with similar appliances and asked them to choose the one that minimized lifetime cost (Blasch et al., 2019). It was varied whether annual energy consumption was displayed in monetary terms or physical units. The correct answer was chosen more often in the monetary label condition. However, this relative difference masked

[4]For example: "A Ford Fiesta FWD will save you $8,689 over its lifetime compared to a Ford Crown Victoria Ffv. That's the same as it would cost for: 17.4 iPads, 8.7 tickets to Hawaii, 174 pairs of Levi's jeans." (Allcott and Knittel, 2019, Figure 1, p. 8)

a wider trend - over half of participants did not report trying to solve any optimisation problem, which speaks to the central role for aspects of bounded rationality.

3.2.3 Concentration Bias and Underestimation Bias

At the core of concentration bias is the idea that larger differences attract more attention that smaller differences. Due to this increased attention, attributes that have larger magnitude differences receive more decision weight. It is important to emphasize that concentration bias is not synonymous with present-focused time preference. For example, the utility value of a large benefit in the future - such as a promotion or sizable bonus - would be diminished by discounting, but attract focus-weighted utility under concentration bias, leading to many small costs on the road to that attention-grabbing prize to be underweighted. However, in the case of pay-now, benefit-over-time investments, concentration bias pushes choices in the same direction as present-bias.

The empirical evidence for concentration bias is mixed. Concentration bias predicted labour supply decisions in a real-effort task (Dertwinkel-Kalt et al., 2022). The effect was mediated by the 'accessibility' of the difference in workload between the two options. However, no evidence for concentration bias was found in a field experiment in Kenya on purchasing an energy efficient heat-stove (Berkouwer and Dean, 2019). More recently, in a consumer choice experiment, a model of relative thinking was found to better fit the data than concentration bias (Somerville, 2022).

Evidence for underestimation bias was recorded in a laboratory experiment and a field experiment in supermarkets (Scheibehenne, 2019). Others experiments have found that it partially explains undervaluation of lotteries relative to expected payoffs, which is typically attributed only to risk aversion (Olschewski et al., 2021). Studies that investigate the related concept of intuitive averaging also record a tendency towards underestimation, which is consistent with a compressive number line (Brezis

et al., 2015). In recent research (presented in the previous chapter), McGowan, Denny and Lunn (2021) found that underestimation bias is present when participants sum sequences that are economically meaningful and familiar, such as utility bills. They also found the bias manifested in a forced-choice task as well as a judgment task where the best guess for the sequence sum was typed.

Underestimation bias and concentration bias make very similar predictions. However, it is still necessary to test their relative predictive power, because each has different implications for policy. Underestimation bias is a clear mistake which drives a wedge between true preferences and choices. Policy interventions that shoulder some of the arithmetic burden for consumers would be valid to correct this 'internality'. For instance, providing lifetime fuel costs instead of monthly fuel costs could be welfare-improving. Concentration bias on the other hand is not a mistake at the time the decision is made. Kőszegi and Szeidl (2013) explain that choices are made based on focus-weighted consumption utility, but these focus-weights do not reflect welfare from experienced utility. A policy maker would need to assure the car seller and buyer that the imposition of lifetime fuel costs is justified by the future interests of the buyer. This is similar to the argument to restrict the temptation available to present-biased individuals. A related difficulty is that because attention is limited, drawing focus to an attribute like lifetime fuel costs could mean other attributes are ignored, or in technical terms, receive lower focus-weight; tackling concentration bias might be akin to pushing down a waterbed. Of course, more drastic policy measures such as enforced sequential decision making, or use of expert third parties, might circumvent the problem of limited attention.

3.2.4 Decision Framework

Following Allcott (2016), assume fully informed and optimizing consumers purchase an energy-efficient upgrade E if and only if: the net benefits, namely the usage utility surplus, v, plus energy cost savings, s, outweigh the relative purchase price:
if v + s > p.

Concentration Bias: Let ß be the concentration bias parameter, with ß < 1. Assume the cost savings are spread over time. A consumer afflicted by concentration bias purchases E if and only if $v + ßs > p$.

Now, reforumlating the benefit-cost trade off so that the additional price is disaggregated over time and the monetary savings are aggregated, the inequality to determine purchase becomes: $v + s > ßp$.

A price generated to match the left hand side will be greater when ß is present, than when it is absent (i.e. equal to 1).

Underestimation Bias

Alternatively, consumers' beliefs about energy costs could also be biased by factor θ, an underestimation parameter that is governed by processes of numerical cognition. We replace the ß with θ. If the cost saving must be accumulated, then the perceived cost-benefit tradeoff that determines purchase is: $v + \theta s > p$
If instead the total purchase price must be accumulated (for instance when it is a monthly instalment on a multi-year plan) then purchase if: $v + s > \theta p$.

If we assume that $\frac{\partial \theta}{\partial length} > 0$, that is, underestimation bias is increasing in the length of the sequence to be summed, we can make predictions about relative underestimation between different framings of costs and benefits for a given E.

3.3 The Experiment

The experiment scenario involved leasing a plug-in hybrid car on a three-year contract. The lease contract took the form of an upfront deposit and 36 monthly payments.[5] The participant was shown two cars, labelled A and B, which were identical except Car B had better fuel efficiency. After viewing the 'personalized projected fuel cost' label for the two cars, participants inputted their additional willingness-to-pay for Car B. Participants did this twice, with two pairs of cars that had a similar difference in fuel costs: the annualised saving was €252 in one choice set, and €324 in the other. The order of presentation was counterbalanced. Two responses were obtained to increase statistical power to detect an effect.

Two features of the fuel cost forecast label were varied across conditions. First, we varied the accessibility of the cost difference. Accessibility is a term for the ease (or effort) with which particular mental contents come to mind: "Highly accessible features influence decisions, whereas features of low accessibility are largely ignored." (Kahneman 2003, pg 11). Dertwinkel-Kalt et al. (2022) found that accessibility mediated concentration bias. To vary accessibility, on half of the labels the cost difference between the cars was implicit. On the other half, the precise difference was given; in other words, the arithmetic was done for participants.

The second manipulation was designed to test a prediction in Kőszegi and Szeidl (2013) that concentrated benefits (fuel savings in this setting) should be preferred to dispersed ones (Kőszegi and Szeidl, 2013, p.65). To test this, we varied the dispersion of fuel saving in the personalized fuel cost forecast. On half the labels the saving was dispersed across all journey types. In the other half it was concentrated in a single journey category. The size of monetary saving was equal across all conditions.

[5]The lease scenario was chosen over a purchase scenario so that different beliefs about how long respondents would own the car would not be a factor in the WTP responses.

Pre-registration of Design and Analysis

The design, hypotheses and analysis plan were preregistered on the Open Science Framework website after ethical approval had been granted by the ethics committee of the Faculty of Arts, Humanities and Social Sciences at the authors' institution, and before data collection began.

The Preregistration Plan outlined in detail the predictions of both mechanisms and the nature of the separating test between them (see 3.1.2. below). The analysis plan stated that the main regression specification would be a linear mixed model. The power analysis, which is important to minimise bias in published findings (Ioannidis et al., 2017) found that a minimum sample size of 65 was needed to detect a medium sized effect (d = 0.5) for a difference in mean willingness-to-pay. Using a non-parametric test, the minimum sample size was 67. We planned for a total sample size of 2000 participants, which equated to 83.33 per arm, which would have provided sufficient power. The final sample size was larger (n = 2368), meaning the minimum power requirements were surpassed.

3.3.1 Design and Hypotheses

Three experimental factors were manipulated between conditions.

(i) The temporal framing of fuel cost information – monthly, annual, or three-year.

(ii) The payment schedule for making the additional extra payment for Car B - monthly, annual, or once-off three-year payment (i.e. additional deposit payment)

(iii) Two elements of how the fuel cost difference was presented on the label were varied, namely the accessibility of the monetary saving (high/low) and the dispersion of savings (yes/no). These label manipulations were orthogonal to (i) and (ii).

Cost-Payment Combinations: Hypotheses

The first hypothesis (**H1**) is that WTP will be decreasing in the temporal aggregation of the payment schedule. The second hypothesis (**H2**) is that WTP will be increasing in the temporal aggregation of the fuel cost frame. In combination, these hypotheses predict that WTP will be higher when the payment schedule is more disaggregated than the fuel cost frame. The three combinations where this is the case are shown to the north-west of the dashed line in Table 3.1. Their respective inverses are across the diagonal. The shorthand in Table 3.1 follows row-column convention: the first letter represents the payment schedule category, the second the fuel cost category. We exclude balanced combinations (e.g. Monthly-Monthly) because that could reduce WTP elicitation to a simple matching task for participants. By imposing imbalance between benefits and costs on the temporal dimension, we allow for different focus-weights (concentration bias) and intuitive summation of the more disaggregated component (underestimation bias) in the process of generating a WTP.

		Fuel Cost Benefit		
		3-Year	Annual	Monthly
Payment Schedule	Monthly	M3	MA	
	Annual	A3		AM
	3-Year		3A	3M

Table 3.1: Six combinations of payment schedule (row) and fuel cost benefit frame (column). The three combinations to the north west of the dashed line are hypothesised to have higher WTPs than their inverses to the south-east.

Separating Test

The separating test between concentration bias and underestimation bias is described below. First a common prediction is described.

Ordering of WTP gaps: Both concentration bias and underestimation bias predict that when the payment schedule is relatively more disaggregated, WTP should be higher. Moreover, both models make the same prediction regarding the ordering of the size of the WTP gaps. The absolute size of the WTP gap between each combination and its inverse can be measured by taking the difference, northwest minus southeast in Table 3.1. Table 3.2 below labels each gap according with the predicted order and the degree of dispersion between the two factors. e.g. on the first row $A36_{gap} = M3_{wtp} - 3M_{wtp}$, and similarly for $B12_{gap}$ and $C3_{gap}$ on the rows beneath.

Label	Payment, Benefit	Payment, Benefit	Label
A36-High	Monthly payment, 3-year benefit **(M3)**	3-yr payment, monthly benefit **(3M)**	A36-Low
B12-High	Monthly-payment, Annual benefit **(MA)**	Annual payment, monthly benefit **(AM)**	B12-Low
C3-High	Annual payment, 3-yr benefit **(A3)**	3yr payment, annual benefit **(3A)**	C3-Low

Table 3.2: Paired combinations of payment-schedule and fuel cost frame

Both concentration bias and accumulation bias predict $A36_{Gap} > B12_{Gap} > C3_{Gap}$. Concentration bias makes this prediction because the focus-weight ratio is greatest for the A36 pair of combinations, smaller for the B12 pair, and smallest for C3 pair. Underestimation bias predicts this order because the strength of the bias is expected to be increasing in the length of the sequence to be summed. The disaggregated component is most granular (longest) in A36 and least granular (shortest) in C3, relative to the other component.

The test to tease apart the predictive power of the mechanisms exploits the fact that the strength of concentration bias depends on a ratio, whereas the strength of underestimation bias depends on an arithmetic difference. Concentration bias predicts that the intermediate WTP gap, $B12_{Gap}$, should be closer in size to $A36_{gap}$, because $36/12 < 12/3$. In contrast, underestimation bias predicts $B12_{Gap}$ should be closer in size to the $C3_{gap}$, because 36-12 > 12-3.

Labelling Hypotheses

Figure 3.1: Dispersed vs. concentrated fuel cost savings in low accessibility condition. Both savings are €252 per year.

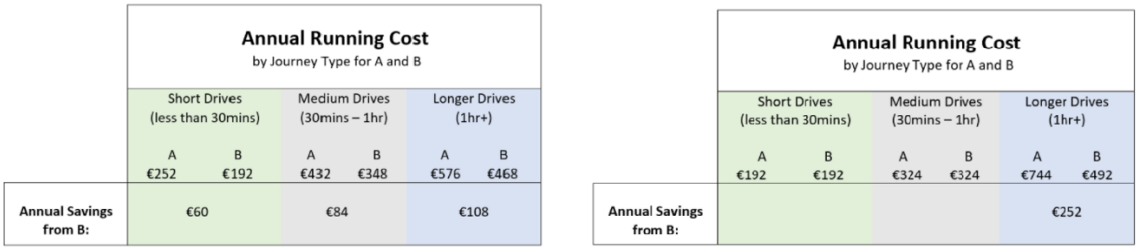

Figure 3.2: Dispersed vs. concentrated savings in high accessibility condition where cost difference is explicit. Both savings are €252 per year.

Figure 3.1 above shows the dispersed saving (left) and concentrated saving (right). Concentration bias predicts that WTP will be higher in the concentrated saving condition. Underestimation bias makes no clear prediction, but note that less arithmetic is required in the concentrated saving frame, which conceivably leaves more cognitive bandwidth for performing intuitive estimation. However, if the total difference in the dispersed condition is estimated by looking at the difference in leading digits, then WTP may be higher in the dispersed condition. Figure 3.2 above shows the same labels except in the high accessibility treatment. If concentration bias is mediated by accessibility (Dertwinkel-Kalt et al., 2022) concentration bias should be reduced in this treatment arm. The arithmetic should also reduce cognitive effort and reduce the effect of left-digit bias on estimating differences.

3.3.2 Method

Participants

The sample was recruited by a market research company to be representative of the population of car buyers in Ireland. The 2,368 participants had all bought a new car in the previous five years or were thinking about buying a new car. Exactly half of the sample was female. The age distribution was as follows: one third were aged 18-34 inclusive, just over two-fifths were aged 35-54, and one quarter were aged 55+. In terms of educational attainment, 53% had at least a higher degree. This is in line with the national average. The household income distribution was also nationally representative.

Each participant was randomly allocated to an experimental condition. There was no significant imbalance across conditions on the main demographic variables (see balance tables in Chapter Supplementary) Participants were paid to take part in this experiment and a preceding survey on features of electric vehicles. The entire survey lasted 10-15 minutes.

Procedure

The online experiment was hosted on the Qualtrics platform. On the first page, the hypothetical scenario was explained: participants were told they were about to lease a plug-in hybrid car on a three-year contract which took the form of an upfront deposit and 36 monthly payments. The choice set comprised two cars, A and B, which were identical except Car B had better fuel economy. Participants were told that they would be asked to input an additional willingness-to-pay for Car B and that "There is no right or wrong answer, but to help you decide, you will be shown personalised running cost forecasts for both cars".

On the second page, the contract for Car A and B was shown, with a question mark

to indicate the additional payment. The deposit was always €4000 and the base monthly payment was €300. Participants were told that after they viewed the fuel cost forecast, they would be asked to input a figure for the question mark to indicate their additional WTP for Car B. After they inputted their WTP, these steps (apart from the introduction screen) were then repeated at the second hypothetical car dealership. For one set of cars, there was a 20 percent difference in fuel efficiency, and for the other pair the difference was 25 percent. The order of the pairs was counterbalanced across participants. Screenshots of each stage of the procedure are shown in the Supplementary Material in Figures 3.6-3.10.

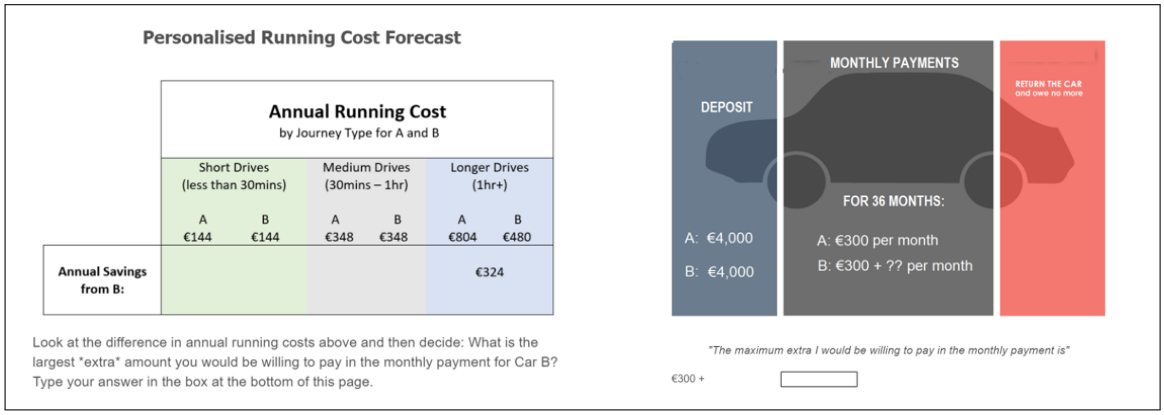

Figure 3.3: WTP figure inputted after looking at fuel forecast

3.4 Results

The results section is structured as follows. First, descriptive results for the six combinations of payment schedule and fuel cost frame are shown in section 3.4.1. In order to allow direct comparison, an annualised willingness-to-pay is calculated from every response: monthly payment responses are multiplied by twelve, deposit-payment responses are divided by three, and annual payment responses left unchanged. Regression analyses are presented in Section 3.4.2.

3.4.1 Fuel Cost-Payment Schedule Results

The annual monetary saving from the greener car was either €252 or €324.[6] Inputting a WTP that exceeds the monetary saving is not a mistake, because one may also value the carbon emissions savings from choosing the more fuel-efficient car, in addition to the monetary savings. However, this argument has a limit: at some point a WTP may be considered 'unreasonably' high. We applied a maximum cutoff at an annualised WTP of €650. This figure was chosen because it is double the maximum annual saving, and thus allows the monetary equivalent of the utility from being environmentally friendly to equal the monetary saving (i.e. v = s in the decision framework terminology of section 2.3).[7]

Table 3.3 below shows the mean WTP for each combination. Recall the rows represent payment schedules, and columns represent fuel cost frames. In order to compare WTP for a given payment schedule, look along that row. To compare WTP for a given fuel cost frame, look down that column. The hypothesised direction is that WTP will be decreasing as one goes from top to bottom along a column, and decreasing as one goes from left to right across a given row.

		Fuel Cost Benefit		
		3-Year	Annual	Monthly
Payment Schedule	Monthly	€300	€313	
	Annual	€214		€174
	3-Year		€169	€139

Table 3.3: Mean Annualised WTP for Each Payment Schedule - Cost Frame Combination. Six combinations of payment schedule (row) and fuel cost benefit frame (column). The three combinations to the north west of the dashed line are hypothesised to have higher WTPs than their inverses to the south-east.

[6] Four of the 48 labels had a typo which meant the actual difference was €312 instead of €324.

[7] The sensitivity of the results to this cutoff is shown visually in Figures 13 and 14 in the Supplementary Material, in line with the preregistration plan.

Table 3.4 below shows the three WTP gaps. The size order of the WTP gaps is also in line with the mutual prediction of underestimation bias and concentration bias. The size of the intermediate $B12_{gap}$ is closer to the largest $A36_{gap}$ than to the smallest $C3_{gap}$: $|161-139| < |139-45|$. This is line with the concentration bias prediction, and the opposite of what underestimation bias predicted.

Label	Payment, Benefit	Payment, Benefit	Label	WTP Gap
A36-High	Monthly payment, 3-year benefit **(M3)**	3-yr payment, monthly benefit **(3M)**	A36-Low	€300 - €139 = **€161**
B12-High	Monthly-payment, Annual benefit **(MA)**	Annual payment, monthly benefit **(AM)**	B12-Low	€313 - €174 = **€139**
C3-High	Annual payment, 3-yr benefit **(A3)**	3yr payment, annual benefit **(3A)**	C3-Low	€214 - €169 = **€45**

Table 3.4: Difference in WTP between combination pairs (A36, B12, C3)

3.4.2 Regression Results

Regression analysis was conducted to test whether the significant differences recorded above were robust to controlling for factors that could influence WTP responses. For example, is the significant difference in WTP across payment schedules and fuel cost frames robust to including label type and demographic factors as controls? We also account for the repeated-measures nature of the data (each participant made two responses).

The first specification in Table 3.5 is a random-intercept linear regression. The dependent variable is the annualised WTP. The independent variables are categorical variables for payment schedule (0 = monthly, 1 = annual, 2 = three-year), fuel cost frame (0 = monthly, 1 = annual, 2 = three-year), and a label type categorical variable for whether the label had dispersed savings across three categories or concentrated savings in one category (0 = Disp, 1 = Con) and whether the cost difference in each category was implicit or explicit (0 = Imp, 1 = Exp). A random intercept allows for random variation in WTP at the participant level. The demographic controls are gender, age category and education (higher degree or not). The trial controls are

dummy variables for specific pair and trial order. These coefficients are not reported in the table to save space but in text we refer to their direction and significance.

	Full sample (WTP ≤ 650)	Zero WTPs omitted
Payment Schedule (ref. = monthly)		
Annual Payment	-113.0***	-166.7***
	(14.52)	(13.71)
Once-off payment	-150.3***	-200.7***
	(14.24)	(13.49)
Fuel cost frame (ref. = monthly)		
Annual fuel cost frame	36.06***	33.76***
	(11.40)	(11.22)
Three-year fuel cost frame	35.22***	40.39***
	(11.32)	(11.21)
Label Type (ref = DISP-IMP)		
DISP-EXP	-7.049	-5.958
	(12.02)	(11.64)
CON-IMP	-10.31	-3.153
	(12.01)	(11.68)
CON-EXP	-32.19***	-11.03
	(12.08)	(11.83)
Constant	344.3***	400.5***
	(19.18)	(18.68)
Observations	3106	2699
Trial and Demographic Controls	Yes	Yes
Zero WTP Responses Included	Yes	No

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

Table 3.5: Random-intercept models for the annualised WTP by condition.

The results in Column 1 align with simple descriptives in Table 3.3. Relative to the monthly payment schedule, the annual and three-year payment schedules record

significantly lower WTPs of €113 and €150 respectively ($p < 0.001$). This is approximately a 33% and 44% decrease relative to the constant of €344. This finding supports Hypothesis 1. The annual and three-year fuel cost frames recorded significantly higher WTPs than the monthly cost frame, with point estimates of €35 and €36 euro respectively. This result broadly supports Hypothesis 2, however note that the effect appears subject to diminishing returns, with no difference between annual and three-year frames.

Regarding the label manipulations, WTP is highest in the label where savings are dispersed and the accessibility of the difference is low (Disp-Imp). WTP is significantly lower in the opposite label, where savings are concentrated and explicit (Con-Exp). Changing the reference category to two other label types (Disp-Exp and Con-Imp) showed that these labels too had significantly higher WTPs than Con-Exp. This indicates a negative interaction effect between the two manipulations: on average WTP was lower for a single, highly accessible difference than when people had to approximate and/or aggregate multiple small differences to generate a WTP.

We omit the zero WTP responses in the Column 2 specification. Zero responses are sometimes interpreted as a protest response (Söderberg and Barton, 2014) but most likely indicate a true preference. The highest proportion of zero responses is in the 3M combination (9.6%)[8]. This combination was hypothesised to have the lowest WTP of all six combinations. Its higher incidence of zeros is consistent with the 'true preference' interpretation. Regardless of the ultimate cause, responses of zero indicate that "experimenter demand" was not strong, increasing one's confidence in the likelihood that positive WTP responses were genuine.[9] The results are broadly similar when one omits zero WTPs. The magnitude of the payment schedule WTP difference becomes larger, increasing to -€166 and -€200 for annual and three-year respectively. The WTP difference by fuel cost frame alters only slightly, with the WTP premium from showing annual cost decreasing from €36 to €33.76, and the

[8]Frequency in the other conditions: 3A (6.8%), AM (8.8%), MA (8.8%), M3 (8.9%),A3 (8.7%)

[9]The zero WTP responses can also be dealt with using a censored Tobit regression. We show results using this approach in Table 3.8 in the Supplementary Material. The main results are unchanged.

premium on the three-year cost frame, relative to monthly, increasing from €35 to €40. The Con-Exp label no longer has a significantly lower WTP, indicating that this label type had recorded a higher number of zero WTP responses.

Some significant demographic differences emerge. Female participants had marginally lower WTP on average (β = -14.6, z = -1.71, p-value = .087) in the first specification, with a β of -15.5 in the second specification (p = 0.062). There was a strong and consistent effect of age category on WTP, with the youngest age category, 18-34, recording significantly higher WTP. The two other cohorts, 35-54 and 55+, had WTPs €40 and €42 lower respectively (p < 0.001) The difference is shown graphically in the Supplementary Material in Figure 3.14. WTP was lower on the second trial (β = -8.35, z = -3.31, p = 0.001) and higher in the scenario where the fuel cost difference was larger (β = 6.61, z = 2.62, p = 0.009).

Predicting WTP Gaps

Table 3.6 shows the difference in WTP across the six combinations of payment schedule and fuel cost frame. The specifications are otherwise the same as Table 3.5. Recall from the design section that concentration bias and underestimation bias made the same predictions regarding the WTP gap ordering ($A36_{gap} > B12_{gap} > C3_{gap}$). But the models differed in whether the intermediate B12 gap should be closest in absolute size to the A36 gap (concentration bias prediction) or the C3 gap (underestimation bias prediction).

The reference category is the A36-high combination, labelled M3, which is shorthand for the combination of a monthly payment and the three-year fuel cost frame. The A36 gap is the difference between the coefficient on M3 (whose value is zero because it is the reference category) and β_{3M} (173.6). The B12 gap is β_{MA} minus β_{AM} (14.7 - (-144.5) = 159), and lastly the C3 gap is β_{A3} minus β_{3A} (-102 - (-146.5) = 44). But a Wald test of coefficient equality shows that the difference between Gap A36 and Gap B12 is not significant (p = 0.58) i.e. the gaps are the same size. Recall that both

models predicted that Gap A36 would be larger than Gap B12. Therefore, although the point estimate suggests the *ratio of imbalance* in concentration bias is the better predictor of WTP gaps, since a prerequisite condition is not met, the separation test between models is inconclusive.

	Full sample (WTP ≤ 650)	Zero WTPs omitted
ref. = Monthly-payment, 3-year cost frame (M3)		
Monthly payment, annual cost frame **(MA)**	14.71	10.06
	(23.56)	(21.5)
Annual payment, 3-year cost frame **(A3)**	-102.1***	-154.0***
	(19.1)	(17.8)
3-year payment, annual cost frame **(3A)**	-146.5***	-203.1***
	(18.7)	(17.50)
Annual payment, monthly cost frame **(AM)**	-144.5***	-202.2***
	(18.9)	(17.75)
3-year payment, monthly cost frame **(3M)**	-173.6***	-226.9***
	(18.8)	(17.75)
Constant	372.4***	432.2***
	(20.8)	(19.6)
Demographic, Label and Trial Controls	Yes	Yes
Zero WTP Responses Included	Yes	No
N	3106	2699

Table 3.6: Random-intercept models for six combinations of payment schedule and fuel cost frame. The three WTP gaps are ref - 3M (A36 gap), MA - AM (B12 gap), and A3 - 3A (C3 gap).

Evidence on the relative predictive power of each mechanism can be considered more holistically by considering a range of WTP cutoffs. Figure 3.4 below shows how the three WTP gaps change in magnitude as the max admissible WTP increases. The A36 gap is shown in red, the B12 gap in blue, and the C3 gap in grey. Both mechanisms predict the red line should lie above the blue line which should lie above the grey line. Looking at Figure 3.4, this predicted order is the exception rather

than the rule. Note that before the €600 responses are admissible, the red line lies above the blue, with the grey line rising steadily to intersect the blue line around the €400 mark. The expected ordering occurs after the €600 responses are included, but only momentarily, with the blue B12 line lying above the red A36 thereafter, with the difference falling short of significance. However, if the required conditions are loosened to A36 ≥ B12 ≥ C3, then the evidence can be interpreted as favouring concentration bias.

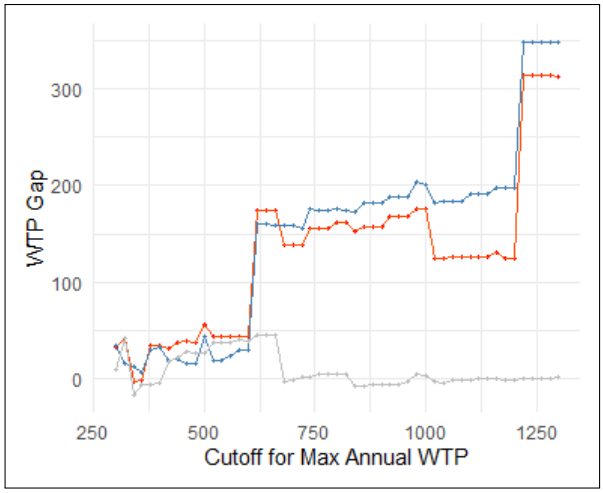

Figure 3.4: Difference in size between three WTP gaps. Red line = A36 gap, Blue line = B12 gap, Grey line = C3 gap. The predicted order is red > blue > grey.

Logistic Regression

Transforming the monthly payment and three-year (once-off) payment to an annualised WTP is necessary to compare the cognitive mechanisms. However the transformation arguably amplifies noise in the responses if people were drawn to the same round numbers within their payment conditions.[10] From another perspective, it is equally valid to simply check if the inputted additional WTP matched or exceeded the monetary saving on offer within a given payment condition. The likelihood of reaching this threshold can then be compared across treatments, without concerns

[10]See Figure 3.11 right panel in Supplementary Material which shows how €50 was disproportionately popular as a response, relative to €30 and €40.

about transformed responses distorting effect sizes. A binary 'WTP threshold' variable was created to indicate whether the inputted response matched the saving per period on offer. A logistic link function was used to regress this binary outcome on the experimental treatments and demographic variables.

The results are shown in Table 3.8 in the Supplementary Material. Relative to the monthly payment condition, the log-odds of a matching response are -1.82 and -1.99 in the annual and three-year payment conditions ($p < 0.0001$). In terms of odds-ratio, these equate to a threshold response being six to seven times more likely in the monthly payment condition. The difference is even larger if zero WTP responses are omitted. For the fuel cost frames, the log-odds of a threshold response is only significantly higher for three-year frame ($\beta = 0.393$, $p < 0.01$) relative to the monthly frame. The results of the labelling manipulations are nuanced. Relative to the reference category of the dispersed-implicit saving label, the concentrated-explicit label (Con-exp) has lower log-odds of a threshold response ($\beta = -0.243$, $p < 0.01$). This accords with the annualised WTP models. However, the dispersed-explicit label (disp-exp) has a higher likelihood of threshold responses ($\beta = 0.275$, $p < 0.05$) with the effect becoming stronger when the zero WTP responses are omitted ($\beta = 0.358$, $p < 0.01$). This result did not appear in the earlier models, where this coefficient was always negative. In terms of demographics, female respondents are less likely to input a threshold response ($p < 0.05$), and age groups of 35-54 and 55+ are substantially less likely to do so (β range: -0.35 to -0.605, $p < 0.01$), with the effect size equivalent to or stronger than the effect of the three-year fuel cost frame. Lastly, a threshold response is less likely on the second trial, and also less likely for the larger difference in fuel costs (€324, or €312 for four of the twenty-four conditions) which perhaps indicates a central tendency in responses.

3.5 Discussion and Conclusion

The 'energy paradox' is one of the most pernicious puzzles in the energy economics literature, and its unresolved nature strongly hints at some pieces being overlooked. Behaviour that appears to be the expression of an individual time preference parameter may have a cognitive antecedent. Allcott (2016) suggested testing concentration bias, which is founded on the concept of limited attention (Kőszegi and Szeidl, 2013). We additionally tested the predictive power of underestimation bias, which is caused by a compressed internal number line (Scheibehenne, 2019). The experimental results show that in the current setting, both mechanisms can explain patterns of WTP not rationalizable under standard time preferences.

Framing fuel costs over a longer period of time increases WTP for fuel efficiency. When people only see the monthly cost, they may underestimate the accumulated fuel savings from choosing the more efficient car. In terms of payment schedules, a monthly payment schedule increases willingness-to-pay relative to the annual payment and once-off additional deposit payment. This suggests that people either (i) underestimate how much they are paying in total when they pay per month and/or (ii) find it less psychologically painful to pay in multiple small instalments. Recall that costs spread out over time receive less utility focus under concentration bias, and this diminishes the pain of paying. The significant difference in WTP across age brackets suggest small payments spread over time may be particularly appealing to younger people, who are accustomed to recurring payments across a range of consumption domains, and may also be more liquidity constrained.

The second objective of the experiment was to explore how WTP is affected by how key information is presented on the label. We imposed two labelling manipulations that were orthogonal to the benefit-payment combinations to investigate this issue. We found WTP is higher when the cost difference is implicit. This suggests that approximating the difference leads to higher estimations on average, possibly due to left-digit bias (Lacetera et al., 2012). This is contrary to the findings of Dertwinkel-

Kalt et al. (2022) who found that explicit differences increased the decision weight given to that component. The second manipulation varied whether the fuel saving was dispersed across three categories of driving or concentrated in the "longer journeys (1hr +)" category. WTP was higher in the dispersed savings category. This may have been due to uncertainty aversion about driving patterns. If the participant thought their driving pattern would differ from the forecast, they would still save money in the dispersed savings condition. In contrast, in the concentrated savings condition, doing relatively fewer 'longer journeys' than forecast would mean a smaller saving. Of course participants may not have internalised the scenario to this extent.

One potential drawback of the experiment is that the key test between the two mechanisms was inconclusive. The pattern of results very tentatively pointed toward concentration bias having higher explanatory power, but this tilt in the balance may have been driven by anchoring effects. Specifically, the base monthly price of €300 potentially anchored both the additional monthly payment and additional annual payment to some extent. This would drag the inputted amounts for these conditions closer together, which means the *annualised* WTP in the monthly payment condition would be much higher. Recall the two combinations that comprised B12-Gap were a monthly payment with an annual fuel cost framing (MA) and an annual payment with monthly fuel cost framing (AM). A low WTP in AM leads to a larger WTP gap (MA - AM) than in the absence of anchoring effects, which tilts the separating test in favour of concentration bias. Future experiments that employ a within-subjects design and a forced-choice elicitation method (instead of producing a number) could shed light on which model contributes relatively more to explaining the 'energy paradox' without caveats.

The lack of a direct measure of time preferences is arguably a limitation of the experiment. However, random assignment to conditions makes it highly unlikely that differences in discount rates were a factor in the results. On the other hand, estimates of the treatment effects may have been more precise had we controlled for this factor. A more nuanced point for consideration is that the division of a time interval can

affect discounting too, a phenomenon called subadditive discounting (Read, 2001; Scholten and Read, 2006). A compressed perception of time has been implicated as a potential cause of subadditive discounting, both by Read and others (Takahashi, 2006). Whether underestimation bias and subadditive discounting have common cognitive foundations is a potentially fruitful topic for research.

Another potential limitation is the hypothetical nature of the willingness-to-pay elicitation, as hypothetical bias can amplify effect sizes (Murphy et al., 2005). This is a minor concern in this experiment for two reasons. First, while people may appear more virtuous when it is free to do so, there is no reason to think hypothetical bias should be vary in strength *between* experimental conditions. Second, the WTP responses of zero, which made up approximately 1 in every 12 responses, indicates low experimenter demand. This should further alleviate concerns about hypothetical bias. More generally, even if one must foreground the caveat of a lack of genuine stakes, hypothetical studies clearly provide important guidance on which interventions are most likely to succeed in changing behaviour.

An issue that transcends the incentivization of the response is whether the findings generalize. This question is always valid, regardless of the strength of an experiment design, or the size of the estimated effect. It is entirely plausible that cohort-specific energy price sensitivity, or local norms regarding making investments that pay for themselves, or beliefs about the social desirability of such actions, would affect the pattern of responses. Answering the generalizability question ultimately requires replication in different settings, which is facilitated by being transparent about the current experiment context and adhering to best practices in open science, both of which we have strived to do. Inductive reasoning about the regularity of the underlying mechanism can also help reduce uncertainty about likely generalizability. To this end, we reiterate that a compressive number line (the basis for underestimation bias), and limited attention (the basis for concentration bias) may be affected by environmental factors, but are not thought to vary intrinsically at the individual level. In contrast, the norm for time preferences is to vary substantially at the individual

level across contexts (Frederick et al., 2002), which perhaps suggests this abstract construct might be a composite measure of preferences and cognitive limitations.

These findings have implication for policy. Policymakers seek ways to make adoption of energy-efficient technologies as painless as possible, both in monetary and psychological terms. Creating opportunities for incremental repayment might be one way to achieve that. However, this issue requires further research, as payment schedules are a dimension of the purchase decision which have received relatively less attention in energy-efficiency research. One reason for this might be an implicit adherence to the idea of descriptive invariance. Alternatively, it may be because payment schedule differences do not fit neatly within the 'information gap' paradigm. To be specific, rational inattention can explain why people might not exert effort sourcing fuel cost information, but it seems implausible for a rational-actor to forego correct integration of total cost once the per-period cost is known. Payment schedules also require more research attention due to the tension in promoting a payment schedule whose effectiveness might stem from inducing the consumer to make a mistake i.e. underestimating how much they would pay in total. But these ethical considerations should be tempered by acknowledging the existing choice architecture in the market: the appeal of modern car financing plans probably comes from inducing similar underestimation or undervaluation.

Similar magnitudes of undervaluation can stem from different channels. Moreover, the time preference channel and the cognition channel are not mutually exclusive. Indeed, if undervaluation is severe, a combination of reinforcing factors might be the most likely explanation. Gaining a detailed understanding of the precise mechanisms underlying what is broadly termed 'myopia' is vital for optimal policy formation. This is especially true when different channels suggest different policy responses. For example, if time preferences are the predominant cause, a regulation that mandates disclosure of annual energy usage would have no effect. However, the same policy would likely be welfare-improving if underestimation bias is the cause of undervaluation, as it would help correct the bias that causes choices to diverge from true

preferences (Beshears et al., 2008).

To conclude, this experiment addresses a gap in the literature by testing the explanatory power of two novel mechanisms that plausibly contribute to the energy-efficiency gap. On an applied level, it is also the first study to simultaneously explore the effect of payment schedule and temporal framing of fuel costs on willingness-to-pay. The energy paradox literature has mostly focused on fuel cost framing, within the paradigm that given the right information, people will make the welfare-maximising choice. In contrast, the marketing literature on how payment structures influence demand has received much less attention (Gourville, 1998). The current findings suggest that offering disaggregated payment schedules may offer the most traction for increasing the uptake of energy-efficient technologies. Results tentatively suggested that concentration bias had relatively better predictive power than underestimation bias. The important point is that both mechanisms, which are not mutually exclusive, had significant predictive power for patterns of WTP not rationalizable under standard discounting. To make progress on solving the energy paradox, the set of behaviours broadly characterised as 'myopia' must be placed under the microscope. To that extent, we hope these findings will help calibrate the lens to bring into focus cognitive foundations that may underlie seemingly paradoxical economic behaviour.

3.6 Supplementary Material

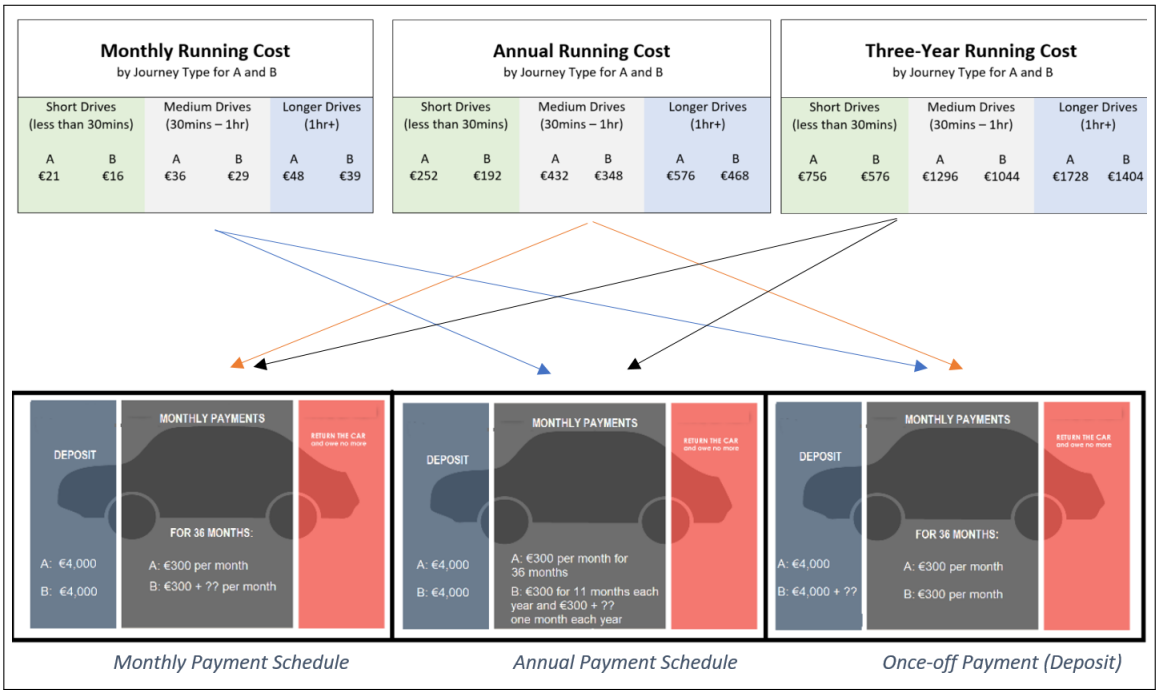

Figure 3.5: Six combinations of fuel cost frame by payment schedule: Top row - temporal framing of fuel cost: monthly (left), annual (centre), three-year (right). Bottom row: different payment schedules: monthly (left), annual (centre), once-off deposit (right).

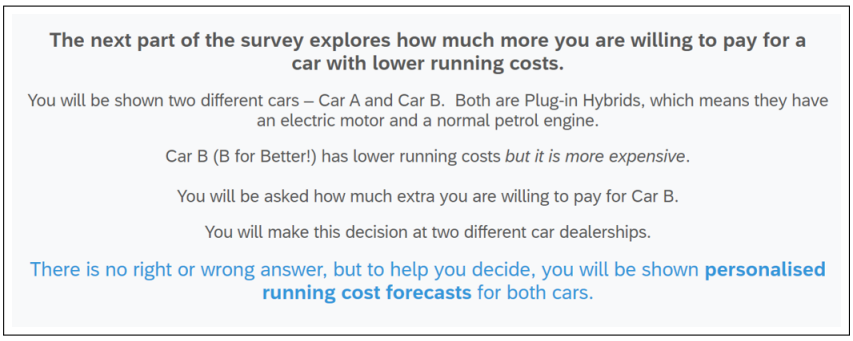

Figure 3.6: Introduction page

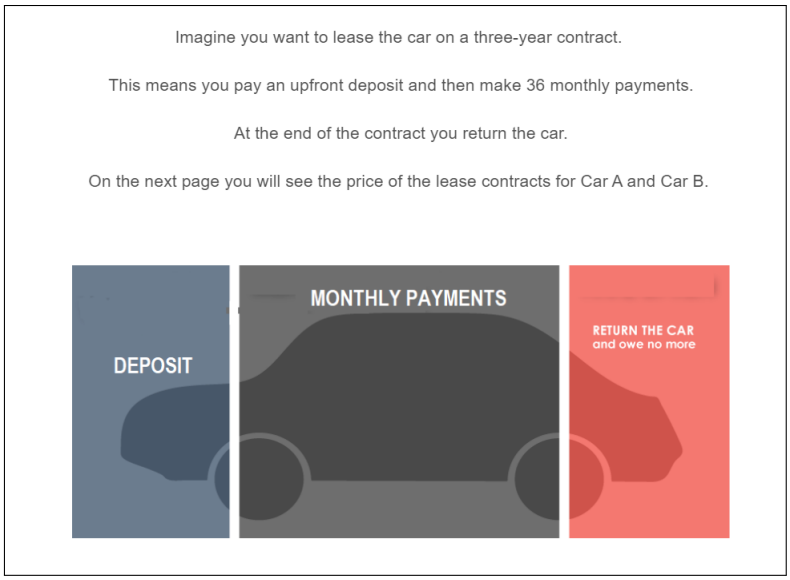

Figure 3.7: Explanation of Lease contract scenario

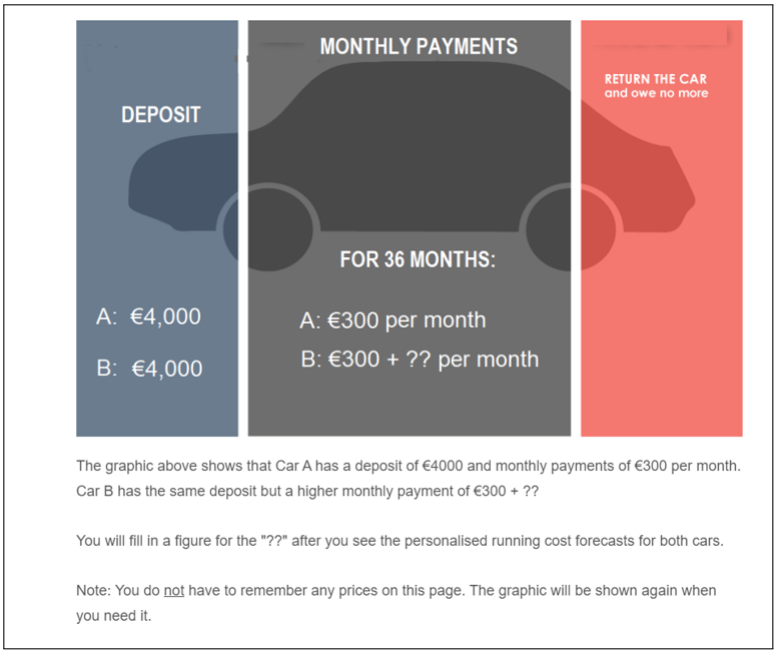

Figure 3.8: Payment Schedule. Current contract with question mark to indicate the contract component for which participants will input an additional WTP.

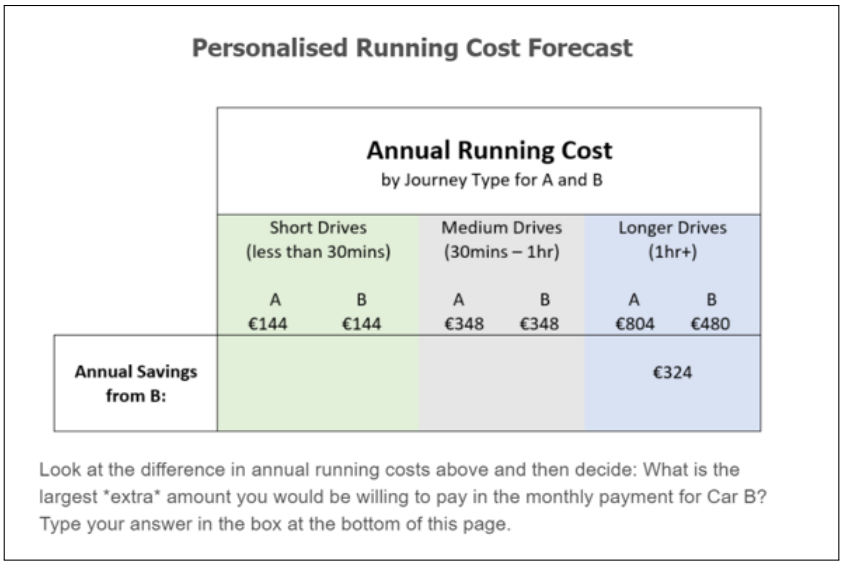

Figure 3.9: Personalised Running Cost Forecast

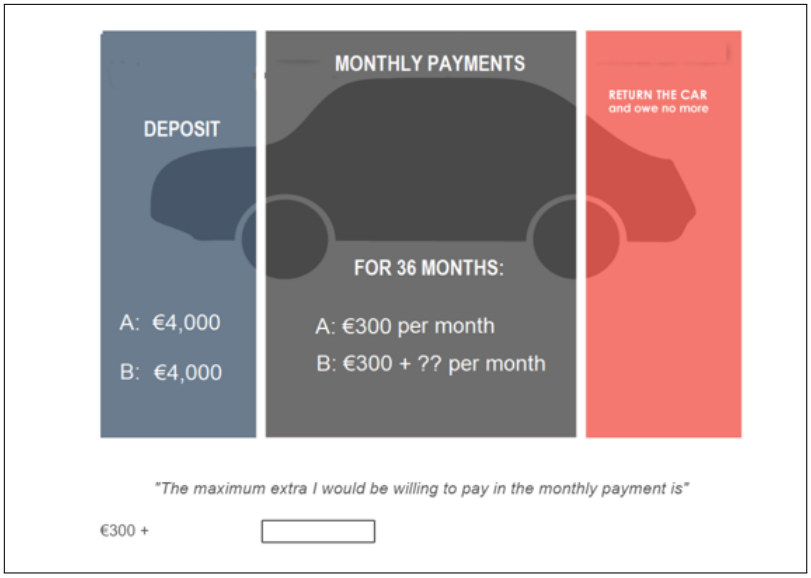

Figure 3.10: WTP figure inputted after looking at fuel forecast (bottom of same page).

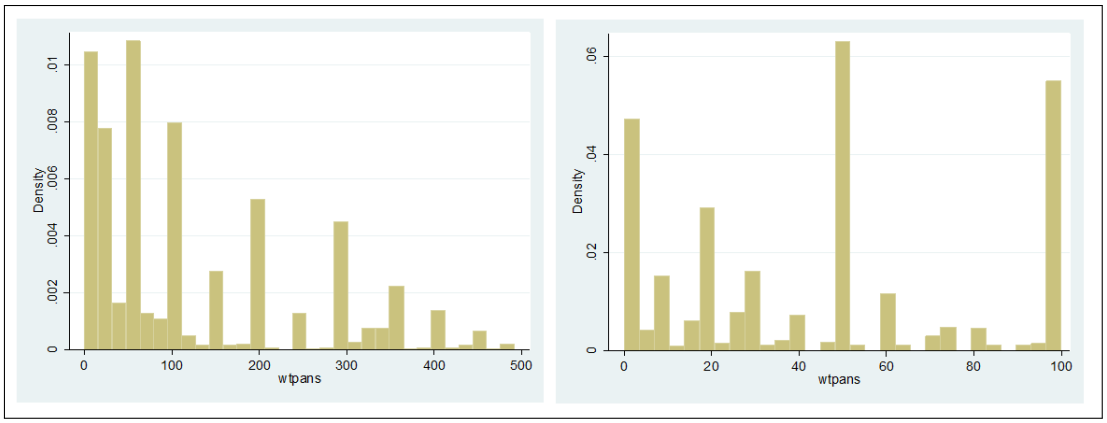

Figure 3.11: Distribution of Monthly WTP. Left panel €0-500. Right panel €0-100.

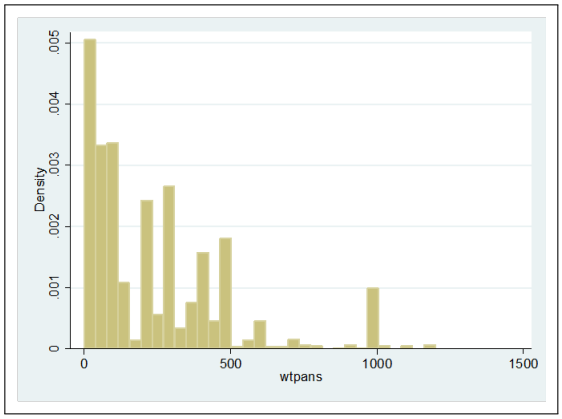

Figure 3.12: Distribution of Annual WTP

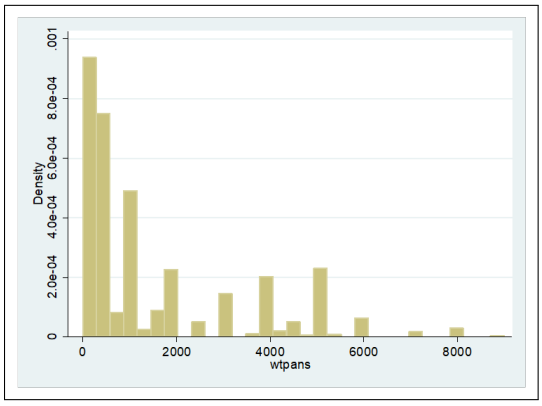

Figure 3.13: Distribution of Three-Year (Deposit) WTP

	All	Zero WTPs omitted
Annual Payment (ref. = monthly)	-101.1***	-159.9***
	(14.46)	(13.78)
3-year payment	-140.2***	-195.8***
	(14.17)	(13.60)
Annual Fuel Cost Frame (ref. = monthly)	35.39***	32.09***
	(11.11)	(10.99)
3-year fuel cost frame	34.28***	40.71***
	(11.18)	(11.16)
DISP-EXP (ref. label = DISP-IMP)	-2.709	-1.811
	(11.80)	(11.49)
CON-IMP	-7.752	0.196
	(11.72)	(11.47)
CON-EXP	-27.21**	-5.913
	(11.81)	(11.72)
Second Trial	-7.522**	-4.075
	(3.171)	(3.231)
Larger cost gap	2.817	3.446
	(3.171)	(3.228)
Female	-12.67	-13.34
	(8.344)	(8.197)
Degree holder	-5.474	-8.780
	(8.454)	(8.297)
Age: 35-54	-38.97***	-30.72***
	(9.589)	(9.423)
Age: 55+	-38.74***	-19.60*
	(11.26)	(11.14)
Constant	325.0***	386.6***
	(19.00)	(18.72)
Observations	3106	2699
Adjusted R^2	0.108	0.197

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

Table 3.7: OLS Results for difference in WTP by payment schedule, cost frame and label type. Standard errors clustered at Participant-level.

	All	Zero WTPs omitted
Annual Payment	-1.819***	-2.224***
	(0.127)	(0.145)
3-Year Payment	-1.994***	-2.384***
	(0.115)	(0.130)
Annual Fuel Cost Frame	0.146	0.0929
	(0.107)	(0.110)
Three-year Fuel Cost frame	0.393***	0.443***
	(0.118)	(0.127)
DISP-EXP	0.275**	0.358***
	(0.119)	(0.127)
CON-IMP	-0.0489	0.0160
	(0.108)	(0.113)
CON-EXP	-0.243**	-0.0643
	(0.110)	(0.115)
Female	-0.178**	-0.185**
	(0.0825)	(0.0872)
Degree holder	-0.0624	-0.0947
	(0.0833)	(0.0882)
Age: 35-54	-0.437***	-0.351***
	(0.0933)	(0.0981)
Age: 55+	-0.605***	-0.428***
	(0.110)	(0.117)
Second Trial	-0.145***	-0.129***
	(0.0445)	(0.0502)
Bigger saving	-0.822***	-0.951***
	(0.0448)	(0.0517)
Constant	2.007***	2.459***
	(0.172)	(0.189)
Observations	4720	4313

Standard errors in parentheses

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

Table 3.8: Logit model for whether inputted WTP was at least as large as monetary saving in given payment frame (e.g. €21 euro per month, €252 per year, €756 over three years).

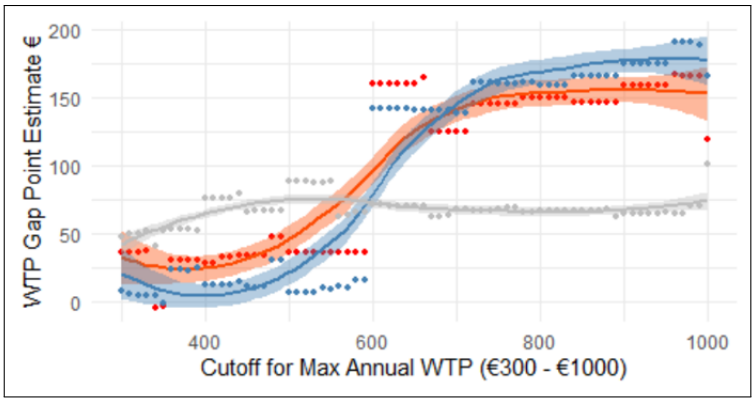

Figure 3.14: Lines of best fit for WTP Gaps as max admissable WTP increases. The predicted order is red > blue > grey.

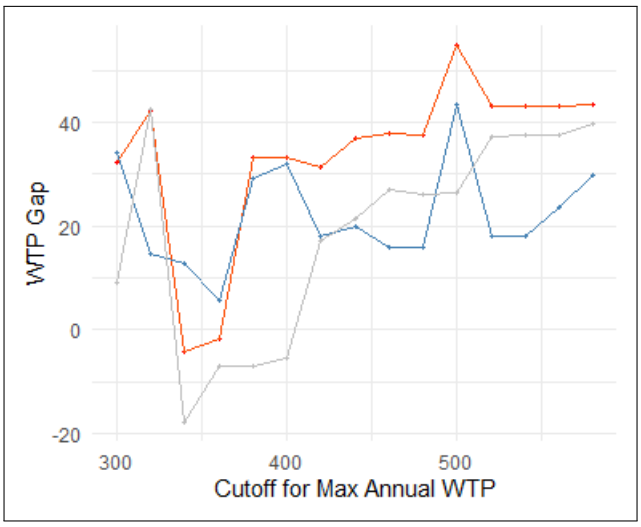

Figure 3.15: WTP gaps below €600. The predicted order is red > blue > grey.

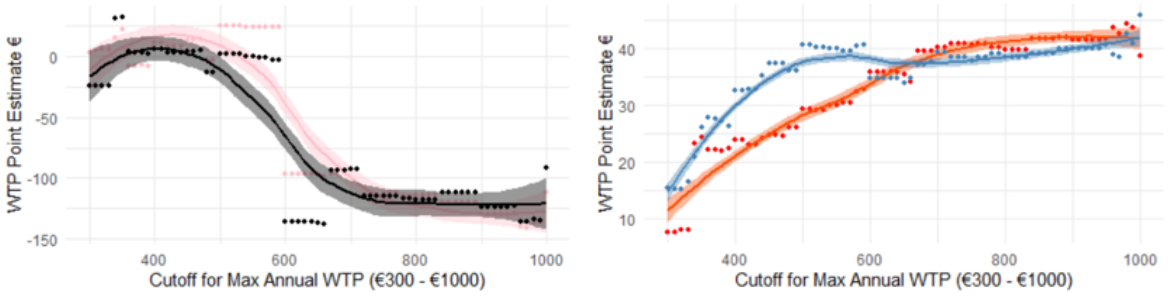

Figure 3.16: Treatment effects across thresholds of WTP. Left panel - difference in WTP of three-year payment (black) and annual payment (pink) relative to monthly payment. Note step change at €600 when €50 monthly repsonses become admissable. Right panel - difference in WTP by fuel cost frame. Three-year frame in blue, annual frame in red. No step change because treatment not sensitive to response format.

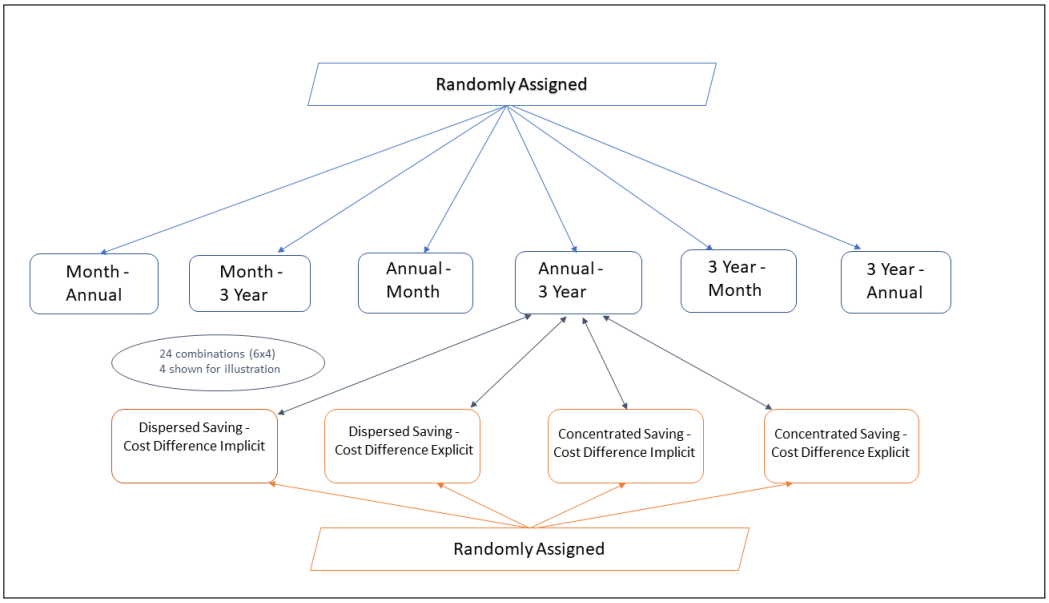

Figure 3.17: Randomisation into conditions: Payment-schedule and temporal framing of fuel cost orthogonal to label type

Saving Frame	Payment Frame	Label Type	Small Saving	Large Saving	Mean WTP (s.e.)
Month	Deposit	Dispersed-NonA	€252	€324	€149 (14.5)
Month	Deposit	Dispersed-A	€252	€312	€142 (13.2)
Month	Deposit	Concentrate-NonA	€252	€324	€144 (11.2)
Month	Deposit	Concentrate-A	€252	€324	€126 (9.9)
Month	Annual	Dispersed-NonA	€252	€324	€167 (12.1)
Month	Annual	Dispersed-A	€252	€312	€176 (13.3)
Month	Annual	Concentrate-NonA	€252	€324	€176 (12.8)
Month	Annual	Concentrate-A	€252	€324	€175 (12.6)
Annual	Deposit	Dispersed-NonA	€252	€324	€187 (11.1)
Annual	Deposit	Dispersed-A	€252	€312	€157 (11.8)
Annual	Deposit	Concentrate-NonA	€252	€324	€177 (12.0)
Annual	Deposit	Concentrate-A	€252	€324	€153 (12.6)
Annual	Monthly	Dispersed-NonA	€252	€324	€355 (26.7)
Annual	Monthly	Dispersed-A	€252	€312	€345 (23.9)
Annual	Monthly	Concentrate-NonA	€252	€324	€291 (25.2)
Annual	Monthly	Concentrate-A	€252	€324	€271 (25.6)
3-Year	Monthly	Dispersed-NonA	€252	€324	€270 (26.2)
3-Year	Monthly	Dispersed-A	€252	€324	€368 (25.1)
3-Year	Monthly	Concentrate-NonA	€252	€324	€333 (26.6)
3-Year	Monthly	Concentrate-A	€252	€324	€229 (23.4)
3-Year	Annual	Dispersed-NonA	€252	€324	€233 (13.4)
3-Year	Annual	Dispersed-A	€252	€324	€204 (11.8)
3-Year	Annual	Concentrate-NonA	€252	€324	€208 (13.4)
3-Year	Annual	Concentrate-A	€252	€324	€212 (14.6)

Table 3.9: Summary of each condition and Mean annualised WTP (sample restricted to Mean WTP between €0 and €650 inclusive).

Chapter 4 Salience, Status and Social Norms: A Natural Experiment on Car Licence Plate Formats

Abstract

People consume particular goods and services to signal a positive social image, also known as status-signalling. Standard models of conspicuous consumption assume that status signals are received by their targets without error. However, limited attention means that subtle signals of status may go unnoticed, or be misperceived. This means salient (i.e. clear and obvious) status signals may be valued more highly. In this paper, I leverage random differences in car licence plate formats to identify the demand for an obviously 'new' car. I use unique panel data of monthly car sales in Great Britain (GB) and Ireland, 2001-2019. Year-of-registration information is shrouded on GB plates, but it is highly salient on Irish plates. A switch from annual to bi-annual plates in Ireland permits a difference-in-differences identification strategy to be used. Results show that the difference in salience has a causal effect on market demand along two dimensions: when purchases occur, and which type of car is purchased. Premium marques such as BMW - conventionally status signals - are less popular when the year-of-registration is more salient, implying substitution between status attributes of marque and age. Overall, these findings add to our understanding of how consumers value salient status signals, and may inform labelling policies to nudge consumption patterns, such as the optimal design of a 'green' licence plate.

This chapter is single-authored. This research was presented at the IEA conference in May 2021. I thank conference attendees for helpful comments.

4.1 Introduction

Despite the common mantra to rid oneself of such concerns, people tend to care what others think of them. Some people take these concerns into the marketplace and consume strategically in order to project a desired social image. Veblen (1899) coined the term "conspicuous consumption" to describe this phenomenon. Spence's (1973) signaling framework gave the behaviour a rational grounding. In Spence's framework, the price of signals is common knowledge. Those who send costly signals reveal they have some desirable quality, like intelligence or wealth. Revelation is in the interests of the signal receiver too. For instance, the employer profits from hiring the most productive worker, so she will examine résumés intently.

It is not obvious that conspicuous consumption offers the same upside to signal receivers. In fact, correctly decoding a status signal might result in envy or other negative emotions. With nothing to gain, signal receivers would be ill-advised to expend scarce cognitive resources decoding signals. This is a problem for the signaller: a status symbol overlooked or misinterpreted equates to money wasted. But correct inference can be almost guaranteed if the signaller can find a symbol that *automatically* captures attention and whose meaning - once attended to - is obvious.

Stimuli that automatically capture attention are known as *salient*. The canonical definition of salience as a phenomenon is as follows: "when one's attention is differentially directed to one portion of the environment rather than to others, the information contained in that portion will receive disproportionate weighing in subsequent judgments" (Taylor and Thompson, 1982, pg 175). Properties that cause a stimulus to be salient include its high contrast with surroundings, it taking an unusual or surprising form, and/or its prominence. Salience can distort economic choices by distorting how decision weights are allocated (Bordalo et al., 2021), and for this reason, salience has been offered as a way to explain why choices diverge from long-term goals. Specifically, the bottom-up attentional system, which is drawn to what is salient (by definition), comes into conflict with the goal-driven, attentional

system, which is also called the top-down system. The bottom-up system often wins (Li and Camerer, 2020). This means a signal receiver, despite her own interests, may be drawn to engage with a salient status symbol. For this reason, it seems theoretically feasible that an increase in the salience of a status attribute will increase its attractiveness to conspicuous consumers.

This paper investigates how a change in salience affects demand for a status symbol. The setting is the car market in Ireland and Great Britain, two countries where there is an age identifier on the licence plate. This information acts as a status symbol because new cars are worth more than old ones, all else equal. The age identifier is highly salient on the Irish plate: it appears as the leftmost piece of information, and begins with the last two digits of the year (e.g. 22 for 2022). In contrast, the age identifier on the British plate is opaque: it is in the middle of the registration format, and a formula is applied which obscures the year the age identifier refers to (details in Section 4.2.3).

This country-level difference in format is not caused by differing consumer preferences for salient year information i.e. the difference is exogenous. The institutional history of licence plate policy illustrates that the addition of a symbol denoting the year of registration in 1960s Great Britain was the solution to the practical problem of a shortage of unused plate combinations. There was no shortage of plate combinations in Northern Ireland, which along with Great Britain makes up the UK. Northern Ireland never added an age identifier because it didn't need to. As consumers in Northern Ireland are similar to those in Great Britain in many respects, the absence of an age identifier supports the claim that its addition in Great Britain was essentially random. The Republic of Ireland changed its plate format in 1987. The impetus was the same capacity issue that Great Britain faced in the 1960s. Ireland's new format gave precedence to the year-of-registration in the licence plate format.

The difference in salience has a causal effect on two dimensions of demand - the timing of car purchases and the type of cars purchased. Regarding timing, in Ireland,

between 22-30% of annual sales occur in each month when a new registration plate is released (henceforth New-Reg month). In contrast, in Great Britain, between 17-22% of sales occur in each New-Reg month. For comparison, in France and Germany, where year of registration is not part of the license plate, monthly sales as a proportion of annual sales *never* exceeded 13% over the period 1990-2019. There is also a timing interaction effect between salience and make type: when the age identifier is highly salient, the proportion of sales in New-Reg months is higher for car makes that would not conventionally be considered status symbols, such as Toyota and Kia. Although the age identifier on a new car is a source of status utility to all purchasers, buyers of Premium makes have an additional source of status in the badge cachet. This means the status utility opportunity cost of delaying purchase is lower for a Premium make.

In relation to the effect of salience on the type of car purchased, the results show that increasing the salience of the age identifier decreases the demand for Premium makes. This implies substitution between status attributes. This causal effect is shown using difference-in-differences analysis which leverages Ireland's switch from annual to biannual plates in 2013. The timing of this switch was not related to economic factors that would plausibly affect car sales, such as consumer confidence or GDP growth. Instead, it was the culmination of a long lobbying effort by the car industry to introduce bi-annual plates (with the hope of achieving smoother sales) which was given extra impetus by the fear a '13' number plate would dampen sales due to superstition over its unlucky qualities. Before 2013, Ireland's premium make market share followed a parallel trend to that of the Great Britain, Northern Ireland and the wider EU, but the trends diverge post-2013. This finding has the most implications for advancing our understanding of how salience effects can alter the nature of consumer demand.

The paper makes three contributions. The primary empirical contribution is testing the effect of salience on consumer choice in a setting where differences in the determinants of salience - contrast, novelty, and prominence - can be isolated. The causal relationship between salience and demand for status attributes is often am-

biguous. Consider for example a classic status good: a Chanel handbag. If Chanel makes handbags larger in a particular season, and sales subsequently increase, is it because the higher salience of the bag - driven by greater prominence - makes it a more attractive status signal? Possibly, but Chanel may have been responding to a consumer preference for larger bags, because they are more practical than small ones and deliver the same instrumental value. It can be difficult to separate signaling motives from direct utility motives. Or, consider that Chanel bags becoming larger could be part of a wider trend in that direction, in which case the larger bag is not more salient relative to comparators. It can be difficult to pin down changes in relative salience in many settings. But these problems do not arise when analysing the relative salience of the age identifier on licence plates.

The second contribution is advancing the literature on numerical biases in the car market. Using German data on used car sales, Englmaier et al. (2018) showed that controlling for the difference in age in months, the price gap in second-hand cars is up to five times larger between December and January registrations, than any other one-month gap. They noted a similar pattern in the mileage dimension, with price discontinuities at 10,000-km odometer marks. This phenomenon is called 'left-digit bias' and has also been recorded in US car prices at the wholesale level, where car dealerships buy cars to sell on to final consumers (Lacetera et al., 2012) and the retail level (Busse et al., 2013). Because all participants at the wholesale level are experienced and transact regularly, their apparent limited attention was attributed to strategic valuation based on what the final consumer would value.

The third contribution relates to understanding the power of labels to shift consumer behaviour (Abeler and Marklein, 2017). From an applied viewpoint, the licence plate may be considered a type of label. The findings underscore that simple labels which highlight salient, socially desirable information can have a substantial impact on consumer behaviour. This is especially pertinent in domains where negative externalities are rife, as is the case for conspicuous consumption, which promotes invidious comparisons (Frank, 1985).

The paper proceeds as follows. Section 4.2 reviews the most relevant papers from the separate literatures on status-signalling and salience, and highlights how the current study begins a synthesis. It also clarifies how age identifiers fit within the salience model of Bordalo et al. (2013) and highlights some limitations of the model as a framework to understand conspicuous consumption. Section 4.3 describes the institutional context and data sources. Section 4.4 describes the results. Section 4.5 discusses these results, considers future directions, and concludes.

4.2 Related Literature

This section situates the contribution of the current paper in the context of the separate literatures on status-signalling and research on salience.

4.2.1 Status Utility: Theory and Empirics

A central tenet of signaling is that low types may try to imitate high types. Several empirical studies document low-income households spending beyond their means in an attempt to 'keep up' with higher income neighbours (Charles et al., 2009; Agarwal et al., 2020; Bertrand and Morse, 2016). Such pooling is impossible if high types send sufficiently expensive signals. Status-seekers may pay more for a good of identical quality because the extra expense sends a signal of status (Bagwell and Bernheim, 1996). In the model of Bagwell and Bernheim (1996), luxury brands are purchased by status-signalers, while others consume "budget" brands, which are cheaper (sold at marginal cost) but identical in quality to luxury brands. The authors note that this pattern appears to manifest in car markets, with some cars being "virtually identical" clones of more expensive models (Bagwell and Bernheim, 1996, pg. 352). This paper also notes that the motive to signal status is tempered by the desire not to been seen to care too much what others think; 'flaunting' wealth can be socially unacceptable.

Cars are a particularly good status symbol because their range of practical benefits allows one a "functional alibi"[1] regarding the fundamental motive for purchase. In the model of fashion cycles developed by Pesendorfer (1995), luxury designs are used as a signaling device in a matching game. A fashion czar (monopolist) sets the design, which is sold to high types who can afford it. Over time the price of the design falls and it spreads across the population. This means it loses its value as a signalling instrument, which suits the monopolist, as high types will now buy a new signal to distinguish themselves once more. Many of the assumptions of this model are contrived and unrealistic (Coelho et al., 2004), but some themes are relevant to the car markets analysed in this paper. The regulator who determines the licence plate format has control over one attribute of the design, much like Pesendorfer's monopolist does. This attribute competes for consumer attention with the attributes designed by car manufacturers.

Natural experiments on status goods are rare, because it is difficult to make a convincing case for exogenous changes to fashion - as noted in the introduction, the changes may be driven by consumer preferences. There are few field experiments too, due to the prohibitive costs involved. An exception is a multi-experiment study on platinum credit cards conducted in conjunction with a bank in Indonesia (Bursztyn et al., 2018). These cards are considered a status symbol because usually only the wealthiest individuals are offered them. Results from a randomised mail-survey experiment showed additional demand for the platinum card, even when its tangible benefits were identical to non-platinum cards. Bank micro data showed that individuals were more likely to use platinum cards in social situations where it would be visible to others, but not in private situations, such as for online purchases.

Using car sales data instead of platinum credit cards to test theories of status utility offers multiple benefits. First, cars are easily visible to people outside of one's immediate circle. Credit cards, by virtue of their size and limited use, are not. Bursztyn et al. (2018) implicitly assume full attention on the part of receivers of the status

[1] This phrase comes from an *Economist* article quoted in Bagwell and Bernheim (1996, pg. 367).

signal. Second, credit card use earns points towards travel or other benefits that are partly monetary substitutes. These factors may partly determine when a credit card is used instead of cash. Such complications do not arise for classic status goods, which are in no way substitutes for cash, nor offer any direct inflow of resources through usage. Third, cars are much more expensive than platinum credit cards and cannot be easily imitated. Lastly, the ordinal rank of status for cars is less context-dependent compared to payment methods such as credit cards versus cash; in some situations, using a credit card might signal higher status than paying in cash, but the opposite might be true too.

4.2.2 Salience

Salience is a feature of an stimulus which grabs attention. Attention can be directed from the top-down, which is called goal-based attention, or from the bottom-up, for example when movement in the corner of your visual field grabs your attention and forces you to look. Schelling introduced the concept of salience in his early experiments on coordination in games (Schelling, 1980). In these experiments, participants tried to match their anonymous partner at a meeting place without communicating. There were literally millions of meeting places in New York City, but the majority of subjects matched successfully by writing down 'Grand Central Station'. Schelling concluded that traditional game theory was deficient in not allowing a role for the salience of decision labels, which he called "focal points". The 'degree' of salience that permits successful coordination was tested by Mehta et al. (1994) in an experiment that distinguished between primary salience, which is what automatically comes to mind, and secondary salience, which is what one thinks will automatically suggest itself to others. If both players execute a secondary salience strategy, they can often succeed in coordinating, even on a location that neither would spontaneously suggest without a coordinating incentive. The experimental results supported the secondary salience mechanism behind successful coordination.

In these coordination games, the decisive salience is the product of the top-down attentional system. The recent literature on the effect of salience in economic choice has focused on the distorting effect of bottom-up salience. Salience can shift economic choices by distorting how decision weights are allocated (Bordalo et al., 2021), and for this reason, salience has been offered as a way to explain why choices diverge from long-term goals. In other words, the bottom-up attentional system, which is drawn to what is salient by definition, comes into conflict with and supercedes the top-down attentional system (Li and Camerer, 2020). The three properties that determine salience in consumer choice are: (i) *contrast* with the attributes of the other options, (ii) *surprise* (also known as novelty), which is judged by the comparison of the attribute value to normal values retrieved from memory, and (iii) the *prominence* with which the attributes are displayed or retrieved (Bordalo et al., 2021).

Before reviewing each property, it is worth noting that a unique aspect of this study is bringing together bottom-up and top-down salience. Only the signalers (i.e. buyers who care about status) are trying to solve a coordination problem. Specifically, they want the receivers' inference to match their intention; they are engaging in secondary salience. However, the payoff of the receiver does not depend on matching. Therefore receivers have no incentive to go beyond using primary salience, which is assumed to be automatic i.e. cognitively costless. This opens the door for attribute salience to determine what receivers attend to, and hence what signalers care about.

Contrast

Contrast between attributes is determined by two factors: ordering and diminishing sensitivity. Ordering simply means that a given attribute is more salient when that attribute value is *more different* from the average value in the choice set. For example, a $1000 price tag for one laptop is more salient if the price of the alternative laptop is $500 compared to when it too is $1000. Diminishing sensitivity reflects the Weber-Fechner law of sensory perception, that a given difference is more noticeable at

lower magnitude levels. For example, a $100 difference will be more salient when the average price is $300 compared to when it is $1000. Ordering and diminishing sensitivity can pull in different directions, a tension which can be resolved by measuring proportional differences between attribute values and choice set averages.

Suprise

To create a measure of surprise it is necessary to model memory. Memory-based salience works like contrast, except values are retrieved. The average of the values retrived from memory becomes the reference point. How retrieval operates is a weakness of the model, but one that is acknowledged and justified on the grounds of tractability. For example, Bordalo et al. (2020) state that similarity is the main driver of associative recall. The following example is given: when asked to name"white things in the kitchen", the cue 'milk' will trigger items such as yoghurt or cheese, which are nearly white. However it seems contextual retrieval might work along a different dimension: 'fridge' could be recalled instead of yoghurt, because it is white and spatially related to milk, despite being less similar in a holistic sense. Another weakness of the memory model is that although Bordalo et al. (2021) discuss databases of memory for different items, they do not discuss the logical schema for ordering observations within this database. If they are stored based on recency alone, it seems it would not be very coherent, and perhaps more difficult to retrieve a sample to compute a norm.

Prominence

Prominence is modelled using Kahnemans's What You See Is All There Is principle (Enke, 2020). Invisible attributes are recalled with probability less than one. This means they are undervalued on average. But unlike contrast and surprise, which arise in stimuli comparison, prominence can be determined by wider contextual factors, in

addition to physical prominence such as size or position. Bordalo et al. (2021) list consumer choice anomalies that they attribute to the prominence channel of salience: the higher incidence of purchasing cold-weather items from catalogues on very cold days, more purchases of convertibles on sunny ones, and higher purchase of health insurance when air pollution is high (Conlin, O'Donoghue, and Vogelsang, 2007 ;Busse et al. 2015; Chang, Huang, and Wang, 2018). The external conditions remind one of positive attributes associated with the items, and interfere with the recall of negative ones. When conditions change, and prominence subsides, the negative attributes come to mind.

4.2.3 Age Identifiers within Salience Framework

This section links the licence plate age identifiers to the three properties that generate salience. The motivation for being explicit about where differences in salience arise comes from the following quote: "A growing body of applied work uses the term salience to explain the outsized role of some information on decisions. But to assess the mechanisms of salience, a researcher needs not just data on information and choices, but also measures of contrast, surprise, and prominence, as well as of sensitivity of choices to these measures" (Bordalo et al., 2021, p. 31)

Figure 4.1 below shows an example of a the current British and Irish licence plates (more institutional details are provided in Section 4.3). There are two differences to note between the plates. These are the position of the age identifier and the format of the age identifier. The positional difference is simple. The British age identifier follows a two-letter area code. The Irish age identifier is the leftmost piece of information on the plate. Regarding the format, the key difference is that the British plate obfuscates the period of registration for six months out of twelve. Moreover, this six month period crosses two calendar years. Specifically, from September to February inclusive, the age identifier is the last two digits of the starting year plus 50. The number 51 in the British licence plate below identifies a car registered between September 2001

and February 2002. From March to August inclusive, the age identifier is the last two digits (e.g. 22 for 2022). The Irish format consists of the last two digits of the year plus a 1 or 2, to represent the half of the *calendar* year in which the car was registered.

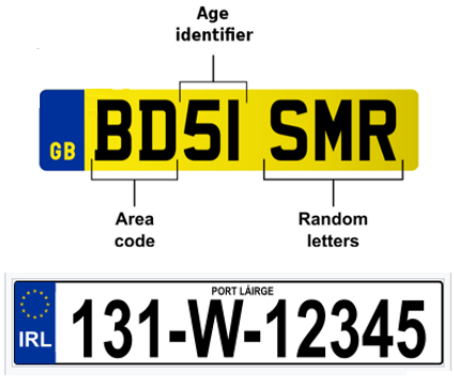

Figure 4.1: Licence plates in Great Britain (top) and Ireland (bottom)

Age Identifiers and Contrast Contrast requires perceiving differences between attribute values. Cognitive limitations do not have a role in this dimension of the BGS model, therefore strictly speaking, the age identifiers do not differ in this dimension of salience. For example, a 2.5 year gap in age, say between a car bought in October 2019 versus April 2017, should generate the same contrast whether it is shown as 192 vs. 171 (Irish plate) or 69 vs. 17 (British plate). But such invariance is unlikely to hold in a richer model of how perceptions of contrast are generated. For instance, adding fifty to half the age identifiers might reduce contrast for a given age gap through the diminishing sensitivity channel of numeric cognition (the compressive internal number line). Also, note that the last digit of the Irish identifier can be ignored and the difference will be correct on average.[2]

Age Identifiers and Novelty: Intuitively, a novel number on a licence plate is surprising because it has not been encountered before. This is not exactly how the

[2] 1.5 year difference for 191 vs. 172 cars, or 2.5 years for 192 and 171. Ignoring the last digit does not bias estimates.

salience model captures surprise. Instead, surprise is determined by the distance between the observed value and the norm in memory. Future work may enrich this aspect of the model, to allow for how surprise-driven salience seems to diminish hyperbolic fashion.

The previous subsection mentioned that a limitation of the surprise component of salience is the highly abstract nature of the recall process. This makes it difficult to define and defend differences in salience on this dimension. To generate a reference point, people need to be able to retrieve a representative sample from memory. This is more likely to happen when there is a logical schema in which the memories can be stored. This seems to be the case in the Irish system, where age identifiers map to calendar years. In contrast, each age identifier in the British system is either 50 higher or 49 lower than the previous one. There is a logic to it, but its opacity means it is harder to reproduce mentally for storage and retrieval.

Age Identifiers and Prominence The Irish age identifier is more prominent for two reasons. First, its position as the leftmost piece of information makes it more prominent than the British identifier, which is in the middle of the registration. The second factor is that wider context of the Irish plate release coinciding with the new calendar year. The "New year, New car" cued association plausibly attracts attention to the age identifier.

4.3 Institutional Context and Data

This section first details the institutional context of changes in the licence plate format in Great Britain and Ireland, and then describes the data sources used in the analysis.

Institutional Context

Great Britain Alphanumeric plates came into effect in the UK in 1904. By the 1960s, due to population growth the British registration system was running out of unused combinations. To fix the problem, the formula was altered and a letter suffix added that denoted the year of registration: 'A' for 1963, 'B' for 1964, and so on. This new system was compulsory for all counties in Great Britain in 1965.

An unintended side effect was that car buyers waited until January to purchase a car in order to get the new plate. Industry lobbied for a change to smooth the seasonal demand this caused. Government acquiesced. In 1967, the 'E' suffix ran only for seven months, from 1 January to 31 July, with the 'F' suffix commencing on 1 August. Henceforth plates were released every August. In 1982, the year suffixes reached Y. Then, the year-identifier was changed to a prefix instead of a suffix, starting again at 'A' for 1983.

British Licence Plate	Time Period
No age identifier	Pre-1963
Age identifier letter prefix, plates released in January	1963-1967
Age identifier letter suffix, plate release in August	1968-1982
Age identifier letter prefix, plate release in August	1983-1998
Age identifier letter prefix, plate release in March and September	1999-2000
Age identifier number code in middle of plate, release in March and September	2001-Present

Table 4.1: Timeline of changes to British licence plate

The January peak that industry complained about in the 1960s had become an August

peak by the 1990s. In 1999, in order to smooth out demand, a bi-annual plate system was introduced, with new plates every March and September. The current registration plate system came into effect in 2001. The registration starts with two letter area code, for example NG for Nottingham. This is followed by the year code. The year code is the last two digits of the year for plates issued between March and August (e.g. '18' for registrations issued between 1 March and 31 August 2018). Plates issued between September and February have 50 added to the last two digits (e.g. '68' for registrations issued between 1 September 2018 and 28 February 2019). One needs to know this simple formula in order to tell what year a car was registered.

Ireland Ireland was part of the UK when the 1904 plates were introduced, and continued to use this system until the late 1980s, when unused combinations became scarce. A new system was introduced in 1987 with simplified county identifiers and a year identifier. The registration began with the last two digits of the year. This was followed by one or two letter combination that identified the county, and with a one- to six-digit sequence number, starting with the first vehicle registered in that area in that year.

In 2013, the age identifier was amended with a 1 or 2 added to signify the half of the year in which the car was registered i.e. 131 for January–June 2013 and 132 for July–December 2013. The official reason for the change was to smooth out sales over the year.[3] The Irish motor trade organisation claim to have been lobbying for five years before the change was made. However, parliamentary records show that concerns that widespread *triskaidekaphobia*, fear of the number 13, would dampen sales gave impetus to the timing of the change (O'Connell, 2012). Thus Ireland followed Britain's example, though without moving the position or format of the age identifier.

[3] "The introduction of the dual registration plate in 2013 was seen as a long-term project for the Motor Industry. The Industry was seeking to shift some activity into the second half of the year to address the serious problem of seasonality in the Motor Industry in Ireland. After many years of campaigning for change a new format was introduced in 2013." (SIMI, 2013).

Plausibly Exogenous?

The multiple changes to the licence plate format in Britain make it unlikely that these changes were driven by consumer preferences. In fact the record is clear that the car industry was taken by surprise at consumers' eagerness to buy a car at the start of the licence plate period. Moreover, the lack of an age identifier in Northern Ireland is further evidence that its presence on the British plate is random; consumers across the UK are alike in many respects. As for Ireland, the most likely reason it adopted a year identifier is because Irish policymakers often copied from their British counterparts.

4.3.1 Data

Data Source	Variable of Interest
1. Motor Trade Organisations in Ireland (SIMI) and Great Britain (SMMT)	Monthly Sales by make (2001-2019)
2. Central Statistics Office	Annual Registration of New and Second-Hand Cars
3. European Automobile Manufacturers Association	Monthly Sales and Make sales
4. Ministry of Transport Test (2005)	Registration of pre-2000 vehicles in Great Britain
5. National Car Test Results (2013-2020)	Reliability of Vehicles by make in Ireland

Table 4.2: Data sources

All data sources are listed in Table 2 above. The primary source is number 1, the monthly sales data from motor organisations. The final panel set comprised all monthly sales of new cars from 2000-2019 for Great Britain, from 2000-2019 excluding 2014 for Northern Ireland, and 2001-2019 for the Republic of Ireland. The 2020 data are available, but not included due to the severe distortion of the pandemic on the timing of sales. The final panel dataset initially consisted of five variables: Year, Month, Jurisdiction, Car Make, and Quantity Sold. Two binary variables were created. One denoted months in which a new registration plate (New-Reg) were released.

The second denoted whether car makes were considered Standard makes (= 0) or Premium makes (= 1).

- *New-Reg*: A binary variable indicating whether a new plate has been issued that month. Equal to 1 for March and September in Great Britain, and for January in Ireland, and also for July (from 2013 onwards).

- *Standard Makes*: The following were defined as Standard Makes, i.e. one would not expect greater-than-average status to accrue from the make badge alone: Toyota, Volkswagen, Honda, Ford, Hyundai, Citroen, Vauxhall, Fiat, Kia, Mazda, Nissan, Peugeot, Renault, Seat, Skoda, Opel, Mitsubishi, Suzuki.

- *Premium Makes*: The following makes were defined as high status brands, also known as luxury or premium brands in the literature: Alfa Romeo, Audi, Aston Martin, BMW, Jaguar, Land Rover, Lexus, Mini, Mercedes-Benz, Porsche, Saab, Volvo.

The primary variables of interest are the proportion of annual sales that occur in each month, the proportion of each car make's sales that occur in New-Reg months, and the market share of different makes. How the raw number of sales varied over the years in the sample is not analysed, but an illustration is provided in Figure 4.10 in the Supplementary Material.

4.4 Results

This section analyses the data to answer three questions, which are listed below.

1. How does the length of the licence plate period affect *when* cars are purchased? (Section 4.4.1)

2. Does the salience of the age identifier have a differential impact on when Premium and Standard makes are purchased? (Section 4.4.2)

3. How does the licence plate format affect *which* cars are purchased? (Section 4.4.3)

4.4.1 Effect of Switching to Bi-annual Plates

Great Britain and Ireland both switched from annual to bi-annual licence plates with the stated policy goal of smoothing the sales distribution over the year, which hereafter will be referred to as 'seasonality'. Whether more frequent plate releases *should* smooth sales depends on the motive for the prior uneven sales distribution. The policy will backfire if people buy early in a licence plate period to maximise the flow of status-utility from having the 'newest' car on the road. With a shorter time period, the depreciation rate on this time-limited flow will increase, which would increase the opportunity cost of delaying purchase. This motive predicts shorter licence plate periods will increase seasonality. On the other hand, smoother sales should transpire if the uneven peak is driven by a *resale motive*: if second-hand buyers base their valuations on coarse information like the year of registration (Englmaier et al., 2018; Lacetera et al., 2012), rather than more precise information like the month of purchase, cars bought at the end of a licence plate period will be undervalued. As the licence plate period shortens, this wedge of potential undervaluation also decreases, and with it the incentive to purchase early. Which of these motives is strongest is an empirical question.

Until 1999, Great Britain released the annual registration plate every August. Figure 4.2 (left panel) shows the proportion of annual sales that occurred in August. Beginning in 1999, plates were released in March and September. This meant the plate released in August 1998 was replaced after only six months. The switch had an immediate and sizeable impact on when car purchases occurred, as shown by the lines indicating the proportion of sales in March and September. Note that sales in

neither bi-annual New-Reg month reached level of August sales.

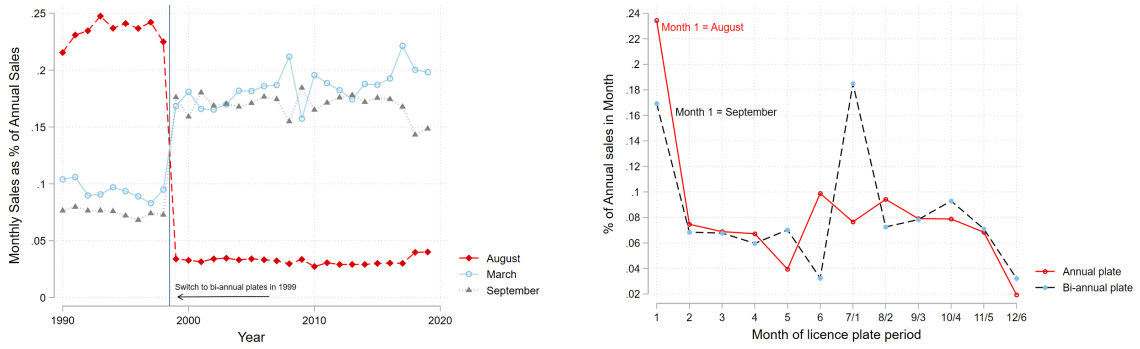

Figure 4.2: Proportion of annual sales in New-Reg months in Great Britain (left). Distribution of sales across the year under annual and bi-annual plates (right).

Ireland introduced a new licence plate with a salient age identifier in 1987. The European sales data begins in 1990, so it is not possible to see the immediate effect of this policy switch on when sales took place. Figure 4.3 below (left panel) shows that the New-Reg month (January) share of annual sales increased gradually. It levelled off in the late 1990s, then increased markedly again from 2000, with the 00 registration. This is unlikely to be a coincidence. It is an empirical regularity in numerical cognition literature that numbers are represented internally on a log scale; subjectively, 1 and 2 seem further apart than 97 and 98 (Dehaene et al., 2008) Recalling the principle of diminishing sensitivity, the *contrast* between the new plate and its predecessor may have increased in the millennium.

Bi-annual plates were introduced in Ireland in 2013, with 131 denoting the first six months and 132 for July onwards. As mentioned in Section 4.3, the policy purpose was to smooth sales over the year. The distribution in the right panel of Figure 4.3 shows this policy backfired. July sales experienced an immediate and rapid increase. The rate of increase was not quite as pronounced as had happened for March and September in Great Britain, where the full adjustment appears to have occurred instantaneously.[4] However, the key point is that January sales did not

[4]This may have been due to a supply constraint, as import data obtained in personal correspondence

decline. Relative to pre-2013, there is a notable decline in sales in April, May and June. We can use the Gini coefficient to measure inequality in the sales distribution. The Gini under annual plates is 0.42, but 0.52 under bi-annual plates.[5]

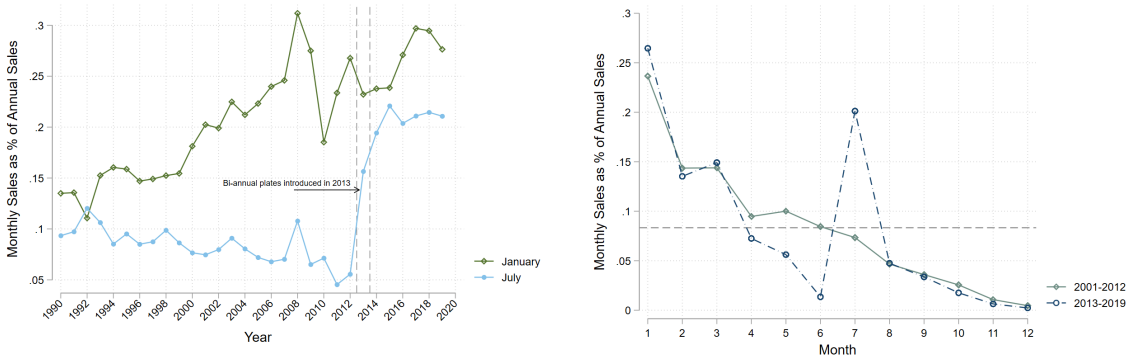

Figure 4.3: Proportion of annual sales in New-Reg months in Ireland (left).Distribution of sales across the year under annual and bi-annual plates (right).

4.4.2 Age Salience and Timing of Purchase by Status Level

In both Ireland and Great Britain, the age identifier causes sales to peak in New-Reg months. This effect is stronger for Ireland. But does the salient year identifier have a differential effect depending on what type of car one is buying? Buyers of Standard makes may have a higher incentive to purchase early, because the age identifier is their only source of status utility. Buyers of Premium makes have two sources of status, the age identifier and the cachet that comes from the marque. This make-specific status is not time-limited, as it is sustained by advertising, cultural norms, and persistent price differences between car makes. This suggests status utility from this source will depreciate more slowly. The implication is that Premium buyers have a lower

from Dublin Port Company show the proportion of new car imports in the May and June 2013 were the same as in 2012, i.e. the industry may not have been fully prepared to meet the seasonal surge in demand.

[5]In contrast, the GB average Gini was 0.27, and for Germany it was only 0.08. Northern Ireland had an average Gini of 0.18 between 2000 and 2010.

opportunity cost of delaying purchase in terms of status-utility foregone.[6] Moreover, independent of status concerns, it is the case that Standard-make buyers are likely to be more conscious of future resale value, because on average their incomes are lower. This is an additional reason why they would be more motivated to purchase at the start of a licence plate period. The previous section noted that at the aggregate level, the resale-motive and the status-motive made different predictions for how the proportion of New-Reg sales would change following the switch to bi-annual plates, however it is the case that at the individual level these motives are reinforcing.

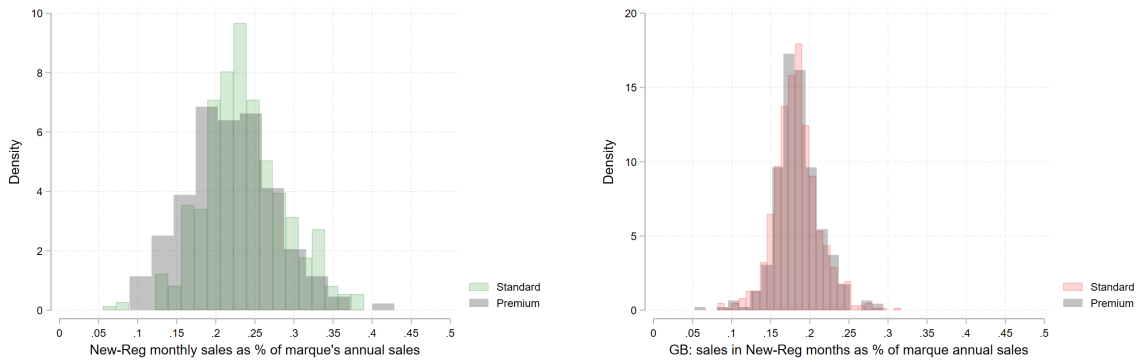

Figure 4.4: Distribution of the proportion of annual make sales that occurred in new-Reg months in Ireland (left). Distribution for Great Britain (right). Premium-make distribution in grey in both.

When the age identifier is the only status source for Standard car buyers, Premium buyers have two sources, it is clear that the difference in motivation to purchase early is increasing in the salience of the age identifier. It follows that Irish buyers of Standard makes may be relatively more incentivized to purchase immediately after a new licence plate is released than British consumers. Results suggest this is the case. In both histograms in Figure 4.4 above, the proportion of Premium make sales that occur in New-Reg months is plotted in grey. On the left panel, note the mass of the distribution for standard make sales in Ireland is shifted to the right (M =

[6] Additionally, delaying purchase ('moving second') may allow uncertainty about the frequency of different makes to be resolved, as they can initially observe which are the most common new cars. This strategy is only feasible when perceiving which cars are brand new is straightforward. And the strategy will only appeal to those who can afford to buy a certain make, conditional on it being less popular.

.236, SD = .0526) compared to Premium makes (M = .218, SD =.0572).[7] different t A two-sample t-test for this difference shows it is significant at the one-percent level (t = 3.63, p = .0002, one-tailed). In contrast, the British histogram shows a near perfect overlay for the proportion of sales for Standard makes (in red, M = .183, SD = .029) and Premium makes (in grey, M = .184, SD = .0286) in New-Reg months.

However, when the age identifier was more salient on the British plate in the 1990s (when the licence plate began with a letter indicating the year of registration) the same pattern of sales in Ireland also occurred in Great Britain. Pre-1999, the proportion of standard makes annual sales that occurred in the New-Reg month (M = 0.241, SD = 0.053) was significantly higher than the proportion of Premium makes registered in this month (M = 0.191, SD = 0.034). An illustration and details of the test are provided in Figure 4.8 in the Supplementary Material.

Regression Analysis

The difference in timing of sales by Make type in Ireland could be driven by make or year effects that are not accounted for in a simple t-test. In Table 3 below, an OLS regression is presented that controls for these factors. The specification is:

$$Y_{m,i,t} = \beta 0 + \beta 1 State_i + \beta 2 Premiummake + \beta 3(State \times Premiummake) + \beta 4 Year + \beta 5 Month + \beta 6 make + \epsilon_{m,i,t} \quad (4.1)$$

Column 1 shows the entire sample of sales proportions in New-Reg months at the make level. Column 2-4 restrict the sample to makes whose overall market share sales was at least 0.5%, 1% and 2% respectively. Focusing on Column 1, the coefficient on the Ireland dummy variable shows that on average, a given car make experiences 7.5

[7] A reason for smoother sales among Premium cars unrelated to any depreciation on the age identifier is that a higher proportion of these cars may be bought by firms for their employees as company cars. The factors that determine when and which company car is purchased are likely to differ from the factors that matter for a private household.

percentage points more of its annual sales in a New-Reg month, compared to in Great Britain. This effect is highly significant (t-stat = 20). The coefficient on Premium make refers to sales in Great Britain only, due to the inclusion of an interaction term. The results indicate that relative to Standard makes, Premium makes experience 2.3 percentage points more of their annual sales in each New-Reg month. This effect is not consistent across specifications: it is highly significant in Columns 1 and 4 ($p < 0.01$), marginally significant in Column 2 ($p = 0.063$) and not significant in Column 3 ($p = 0.37$). The effect size also drops to 1.5 percentage points by Column 4.

In contrast, the coefficient on the $\beta 3$ interaction term, Ireland × Premium make, is highly significant and consistent. The point estimate varies only slightly across sample restrictions, being highest in Column 1 (β = -2.156, $p < 0.0001$) and smallest in Column 3 (β = -2.156, $p < 0.0001$). The interpretation is that Premium makes in Ireland register two percentage points fewer of their annual sales in New-Reg months, relative to Standard makes in Ireland.

	All	≥ 0.5% MS	≥ 1% MS	≥ 2% MS
Ireland (ref = GB)	7.466***	7.397***	7.611***	8.285***
	(0.373)	(0.290)	(0.298)	(0.332)
Premium make	2.331***	2.945*	0.596	1.554***
	(0.859)	(1.583)	(0.665)	(0.498)
Ireland × Premium make	-2.156***	-1.949***	-1.900***	-2.039***
	(0.579)	(0.467)	(0.491)	(0.563)
Constant	14.32***	14.53***	14.27***	13.88***
	(0.488)	(0.441)	(0.462)	(0.503)
Observations	1946	1602	1392	1014
Year, Month, make Fixed Effects	Yes	Yes	Yes	Yes
Adjusted R^2	0.299	0.470	0.516	0.572

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

Table 4.3: Modeling the proportion of make annual sales in New-Reg months.

Note that the standard Ordinary Least-Squares assumption of an unbounded dependent variable is violated when modelling differences in a proportion, which by definition must vary between 0 and 1. To account for this, censored Tobit regressions

for the specification above were conducted. The results from the Tobit models are presented in Table 4.7 in the Supplementary Material. The results are unchanged from the analysis above.

Market-Segment Level Analysis

The unit of observation in the analyses above is the sales proportions at the Make level. The use of percentages gives equal weight to makes with different levels of sales. Restricting the sample based on market share, as in Col 2-4 of Table 4.3, partially solves this minor problem (it would be a major confound if there was reason to believe that car makes with smaller market shares *consistently* have their annual sales concentrated in New-Reg months.) Nevertheless, one solution is to analyze sales at the aggregate market-segment level (Standard and Premium) instead.

The regressions below in Table 4.4 model when sales occur for Premium and Standard market segments. The dummy variables are identical to the specifications in Table 4.3. However, sales proportions are not restricted to New-Reg months; the entire year of sales is included along with a dummy variable to denote 'New-Reg' months. Year and month fixed effects are included as before, but make fixed effects are not possible due to the aggregation.

Results of the simple model in Col. 1 show that New-Reg months in GB experience an extra eight percentage points of annual sales compared to other months. The interaction New-Reg × Ireland shows this effect is twice as strong for Ireland, with an *additional* eight percentage points of sales. The key result is the three-way interaction, New-Reg × Ireland × Premium make (PM), which has a negative coefficient of -0.0284 ($p = 0.019$). This indicates that relative to Great Britain, and accounting for the average difference in propensity to purchase in a New-Reg month between countries, the Irish Premium make segment has 2.8 percentage point lower sales in New-Reg months, relative to Standard-makes in Ireland. This is larger than the equivalent estimate from Table 4.3 using make-level data. In the second column, a 'Bi-annual'

dummy is added to capture the shift to bi-annual plates in Ireland in 2013. The coefficient on the key three-way interaction New-Reg × Ireland × Premium make (PM) falls slightly to -0.0261 (p = 0.029). In the third column, Northern Ireland is added as a control and as a sense-check. Recall that NI registration plates have no age information, so New-Reg is always equal to zero. Note that the coefficient on the baseline NI variable is significant at the 1% level, which reflects its more uniform sales pattern across the year, due to the absence of "New-Reg' months.

	Simple		Ireland Bi-annual		Add NI	
New-Reg	0.0829***	(0.00574)	0.0772***	(0.00674)	0.0704***	(0.00568)
Ireland (ref. = GB)	-0.00555*	(0.00293)	-0.00556*	(0.00293)	-0.00551*	(0.00296)
New-Reg × Ireland	0.0869***	(0.0101)	0.0844***	(0.0101)	0.0808***	(0.00887)
Premium make (ref = Standard)	-0.000836	(0.00267)	-0.00214	(0.00309)	-0.00172	(0.00298)
New-Reg × PM	0.00501	(0.00662)	0.0115	(0.00820)	0.0111	(0.00727)
Ireland × PM	0.00350	(0.00406)	0.00358	(0.00408)	0.00356	(0.00424)
New-Reg × Ireland × PM	-0.0284**	(0.0121)	-0.0261**	(0.0119)	-0.0261**	(0.0109)
Bi-annual Plates Ireland (ref = pre-2013)			-0.00932	(0.00605)	-0.00598	(0.00458)
Bi-annual × New-Reg			0.0156*	(0.00810)	0.0153**	(0.00713)
Bi-annual × PM			0.00355	(0.00444)	0.00240	(0.00334)
Bi-annual × New-Reg × PM			-0.0176	(0.0114)	-0.0165*	(0.00982)
Northern Ireland (NI)					0.0185***	(0.00221)
NI × PM					0.000956	(0.00331)
Constant	0.0802***	(0.00484)	0.0819***	(0.00487)	0.0928***	(0.00446)
Observations	912		912		1308	
Year and Month Fixed Effects	Yes		Yes		Yes	
Adjusted R^2	0.768		0.769		0.755	

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

Table 4.4: OLS Regression of the market-segment monthly sales share.

4.4.3 Salience of Age and Demand for Premium Makes

Does the licence plate format cause demand substitution across classes of car make? There are two reasons why the change to bi-annual plates in Ireland in 2013 might decrease the market share of premium cars. The first is that the shift makes the age attribute more culturally prominent (because the media discuss the reasons for, and potential consequences of, the change). The change in format also increases novelty due to the new pattern. Increases in prominence and novelty make the age identifier more salient, which can increase the decision weight placed on this component. Because salience is a zero-sum game, greater salience on the age identifier lowers the probability of attending to the marque 'badge' attribute. Alternatively, one could think of the 'badge' attribute becoming noisier.

The second reason relates to complementarity between status attributes of age and badge caché. Both absolute age and relative newness may matter for status. Changing from an annual to bi-annual licence plate regime clearly does not change how old a car is, but it does alter how often its rank on 'newness' changes. Compared to under an annual plate regime, a car's rank on ' newness' declines twice as fast under bi-annual plates. If the badge attribute and age identifier are complements in the status utility production function, the higher depreciation rate on age will drag down the status value of the brand itself. This matters more for Premium makes, because they are the ones for whom the store of status value is primarily in the badge. Section 4.6.2 in the Chapter Supplementary Material presents one functional form for how a shortened licence plate period could affect depreciation. The depreciation rate is defined as the reciprocal of the rank. This is somewhat arbitrary but it captures the pattern of an initial rapid decline that approaches zero asymptotically ($1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ etc.) The intuition for how this affects purchase decisions is as follows: if a prospective buyer of a new Premium car foresees this more rapid depreciation (compared to under to the annual licence plate counterfactual), they may be inclined to skip the period of precipitous depreciation and go straight to buying a second-hand Premium

car, or perhaps purchase a brand new Standard one instead.

With these factors in mind, a difference-in-differences (DiD) empirical strategy is employed to estimate the causal effect of the switch to bi-annual licence plates on the demand for new premium cars in Ireland. The key identifying assumption is that the market share of premium makes in Ireland would have followed a parallel trend to GB in the counterfactual scenario where the licence plate did not change to bi-annual. The timing of the policy was random with respect to factors theoretically connected to car sales, such as consumer confidence and GDP growth (see Section 4.3 for reasons for the change). For brevity, the switch to bi-annual plates will be referred to as 'the treatment' hereafter.

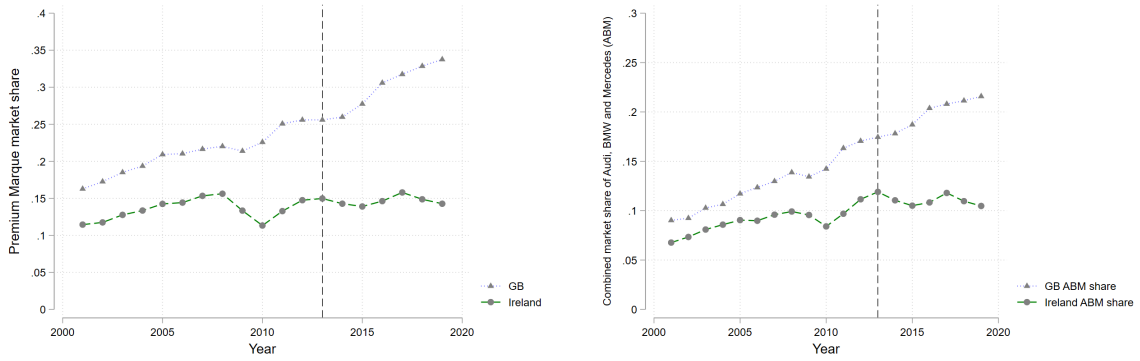

Figure 4.5: Left panel: Premium make market share for GB (dotted-triangle) and Ireland (dash-circle). Right panel: Cumulative market share for three most popular premium makes - Audi, BMW, and Mercedes (ABM).

The trend line for the Premium-make market share is shown in the left panel of Figure 4.5 above. The share in Ireland is lower by a consistent margin. The trends diverge around the time of the financial crisis (2008) which had a bigger impact on Ireland. The gap is wider but stable until the treatment, at which point the trends diverge again. The right panel of Figure 4.5 shows the trend lines for the cumulative market share of the three most popular premium makes, Audi, BMW and Mercedes-Benz (ABM). The divergence is slightly more pronounced. The regression analysis below uses the ABM category as a robustness check on the results.

Alternatively Northern Ireland can be used as a control.[8] Figure 4.6 below replaces GB with NI as comparator to Irish premium make market shares. The trend lines are closer together. There is a similar temporary divergence in trends around the financial crisis. Importantly, a perceptible divergence appears post-treatment, especially in the right panel which uses the ABM sample.

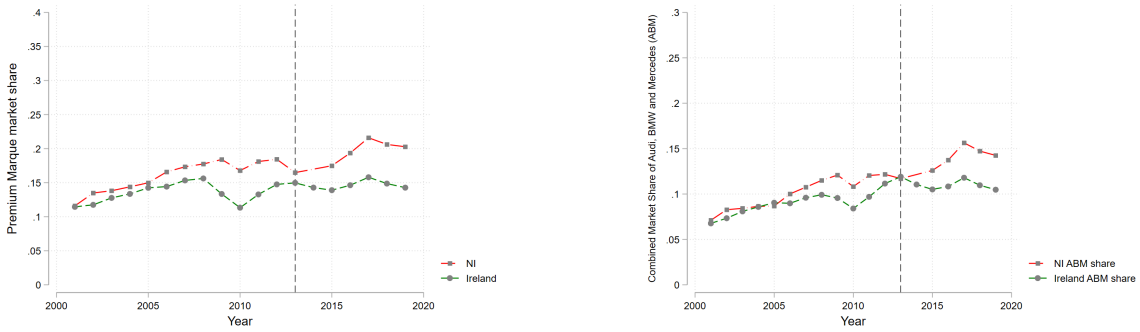

Figure 4.6: Left panel: Premium make market share for Northern Ireland (dashed-square) and Ireland (dash-circle). Right panel: Cumulative market share for three most popular premium makes - Audi, BMW, and Mercedes (ABM).

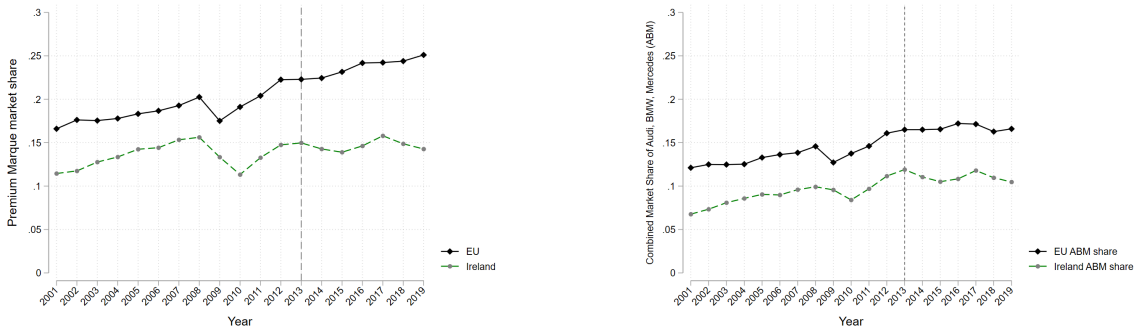

Figure 4.7: Left panel: Premium make market share for EU (black line with squares) and Ireland (green dash-circle). Right panel: Cumulative market share for three most popular premium makes - Audi, BMW, and Mercedes (ABM).

The last comparator is the entire EU car market. Figure 4.7 below shows the trend in Premium-make market share. The pattern here is slightly different. At the aggregated

[8] In section 4.3 it was noted there is some missing data for Northern Ireland. This includes all of 2014, and one quarter of 2009-11, and 2015-17. Car make sales as a percentage of annual sales can still be estimated, but need to be caveated. Details in Supplementary Material.

EU level, Premium make share did not increase to the same extent as in Great Britain alone. However there is still a divergence between the Irish share and the EU relative to the counterfactual. The next section uses a difference-in-differences analysis to determine the significance of this effect.

Regression Analysis

As in section 4.4.2, the first dependent variable used is the make-level market share. The model specification is as follows:

$$Y_{m,i,t} = \beta 0 + \beta 1 TreatedState + \beta 2 PostTreatment + \beta 3 PostTreatment \times TreatedState + \beta 4 make + \beta 5 Year + \epsilon_{m,i,t} \quad (4.2)$$

$Y_{i,m,t}$ is the market share of make m, in state i, at time t. *TreatedState* is a dummy variable that takes the value of 1 for Ireland and 0 for GB (columns 1 and 2) and NI (columns 3 and 4). *PostTreatment* is a dummy variable that takes the value 1 for each year from 2013-2019, and the value 0 for years prior to 2013. The coefficients on $\beta 2$, $\beta 3$ and $\beta 4$ capture make and year fixed effects respectively and ϵ captures idiosyncratic error.

The coefficient on the key interaction term, Ireland × Post-Treatment, is significant in all four specifications. In Column 1, the interpretation of the coefficient is that the average Premium-make experienced a 0.68 percentage point decline in market share due to the treatment. In the restricted ABM sample, the effect size is larger, with the point estimate a 1.65 percentage point drop in market share. This equates to approximately a 33% drop, because the market share of these makes varies between 3 and 5% of overall sales in Ireland. Columns 3 and 4 use Northern Ireland as the control. The difference-in-differences coefficient is still significant at the five percent level, but notably reduced in magnitude.

	GB	GB ABM	NI	NI ABM
Ireland	-0.626***	-1.223***	-0.211***	-0.372**
	(0.0701)	(0.153)	(0.0568)	(0.157)
Post-Treatment	1.217***	3.537***	0.672***	2.071***
	(0.310)	(0.512)	(0.193)	(0.370)
Ireland × Post-Treatment	-0.689***	-1.650***	-0.263**	-0.526**
	(0.152)	(0.277)	(0.105)	(0.257)
Constant	0.231	3.300***	0.0911	2.549***
	(0.222)	(0.365)	(0.177)	(0.336)
Observations	447	114	445	111
Year and make FEs	Yes	Yes	Yes	Yes
Adjusted R^2	0.849	0.795	0.890	0.588

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

Table 4.5: Regression of average market share of Premium makes: Difference-in-differences analysis

Market Level Analysis

In this section, the treatment-control comparison is made at the aggregate market share level. In simple terms, we test whether the visual difference in Figures 4.5 and 4.6 are significant, ensuring that the results in Table 4.5 are not spurious.

The results of this aggregated analysis are consistent with the previous make-level one. In Column 1 (GB) the interaction indicates that the premium-make market share in Ireland was reduced by 7.5 percentage points as a result of the treatment. As a sense check, note that the coefficient of -4.95 in Column 2 divided by 3 perfectly matches the make-level coefficient (1.65) in Column 2 of Table 4.5, as it should. Turning to Column 3 and 4, the difference between Northern Ireland and Ireland is smaller, with an average point estimate of just under two percentage points, but still significant at the five percent level.

A valid concern is that the results might be driven by changes in the exchange rate. Due to the proximity of Ireland to the UK, when the Euro is strong against the pound sterling, second-hand imports usually increase. Assuming the hassle costs of

	GB	GB ABM	NI	NI ABM
Ireland	-7.503***	-3.670***	-2.503***	-0.879
	(0.708)	(0.480)	(0.546)	(0.637)
Post-Treatment	13.94***	10.61***	8.626***	3.970***
	(2.811)	(1.572)	(0.864)	(0.777)
Ireland × Post-Treatment	-7.579***	-4.951***	-2.060**	-1.822**
	(1.386)	(0.864)	(0.897)	(0.878)
Constant	17.61***	9.730***	10.93	9.807***
	(1.458)	(0.786)	(.)	(0.533)
Observations	38	38	38	38
Year FEs	Yes	Yes	Yes	Yes
Adjusted R^2	0.910	0.911	0.775	0.554

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

Table 4.6: Regression Analysis: Dependent variable is Premium make market share.

importing a car are approximately constant across car type, the saving is increasing in the domestic price of the car. In line with this, the Premium share of imports (24% approx) is higher than the corresponding share of domestic new car sales. However, there was no shift in the second-hand market in 2013 that could explain the results. In fact, the premium-share of imports fell in 2013 despite the euro becoming stronger vis-a-vis sterling. This rules out the treatment effect being driven by substitution to imports. Details of the import market are shown in Figures 4.9 and 4.10 in the Supplementary Material.

4.5 Discussion and Conclusion

This is the first paper to apply salience as defined by Bordalo et al. (2021) to conspicuous consumption markets. This domain is one where the salience models could gain much traction, because a prerequisite for successful signaling is that peers attend to the signal.

Taken together, the results show that the format of age information on a vehicle

registration plate, and the length of the licence plate cycle, have a substantial effect on when people buy cars, and also the type of car purchased. It is hard to rationalize the difference in timing for sales for Standard and Premium car makes under a highly salient age identifier without appealing to status motives. Salience as a driver of the timing difference is supported by the similar pattern in Britain when the age identifier was a letter prefix. When the age identifier was made more opaque, being placed in the middle of the format and shrouded behind a formula, the difference in the timing of sales between Premium and Standard makes disappeared.

The most simple, and therefore preferred, explanation is that buyers of Standard makes receive status utility from the Age attribute only. The opportunity cost of delaying purchase is higher for these buyers than for customers of Premium makes, who derive status utility from two sources, the marque badge and the age identifier. Moreover, status utility from make depreciates more slowly (as it is not time limited). But there are other interpretations for the timing difference in sales. One concerns self-signaling and motivated reasoning (Epley and Gilovich, 2016). This is the idea that people generally reason their way to conclusions they favour. People generally prefer to think of themselves as trend setters rather followers. Immediate purchase permits oneself to think others who subsequently buy new cars are copycats. This might, at the margin, be an additional source of utility that compels early purchase. Another complementary mechanism to signaling is mimetic dominance, the idea that private consumption utility is increasing in the exclusivity of an good (Imas and Madarász, 2020).

The switch from annual to bi-annual plates had a negative effect on the market share of Premium makes in Ireland, relative to Great Britain and Northern Ireland. The effect size is substantial. A back-of-the-envelope calculation suggests the revenue foregone to be in the region of €1.8bn.[9] It is intuitive that sellers of Premium makes do not desire other attributes, over which they have no advantage, to crowd-out their

[9]Inputs: Loss of 5% market share per year over 7 years = 35% cumulative. 130k new cars sold per year on average 2013-2019. 130k × 0.35 = 45,500. Multiply 45,500 by price of typical premium car (say €40k approx). 45,500 × €40,000 = €1.8bn approx.

unique selling point as status signals. It diminishes their exclusivity on the dimension on which all cars are similar. Complementarity between age identifier and badge may also degrade the value of the latter. Consistent with this, major car producing countries, such as Germany, Japan, USA and France, do not have any year information on their licence plates.

There are several feasible extensions to the current analysis of the causal effect of bi-annual plates on Premium make market share. The most obvious is to include price data from the second-hand market. One potential reason for the decline in Premium car share from 2013 was that the bi-annual plate accelerated the rate of depreciation on the car, because the age-rank falling faster dragged down the value of the marque cachet (which would be almost timeless if no age identifier was present). Second-hand car sales data would shed light on whether the biannual plate actually had a differential impact on the depreciation rate of Standard and Premium marques.

The analysis presented can also be developed by inclusion of additional control variables, such as tax rates on engine size (premium makes have larger engines on average), taxes on emission levels, and macroeconomic variables such as exchange rates and GDP growth, would reduce uncertainty about the causal effect. Another way to enrich the data is to incorporate model-level data. The marginal Premium-make buyer, who buys the entry level model, is more likely to switch than the buyer who purchases the flagship premium car. With model-level data one could gain a more detailed picture, to see whether actual substitution across makes matches this intuition. Model level data would also allow a more refined stratification of Standard vs. Premium based on the cost of specific car models. Gathering data from more countries would also improve the quality of the inference and may permit complementary identification methods to be used, such as synthetic control (Abadie et al., 2010).

A limitation of this paper is that the supply side response is not examined. The Irish motor industry lobbied for bi-annual plates on the basis of wanting smoother sales.

However, the self-interest of individual dealerships may have sustained advertising campaigns in the original New-Reg month (January). The coordination problem here is obvious. Supply side factors may explain why smoother sales did not occur. But the decrease in Premium-make market share is harder to explain from the supply-side.

The aggregate welfare effect of a salient year identifier is an important question for future research. Answering this requires making assumptions regarding intrinsic quality differences between make types. At one extreme is a Veblen effect, where richer buyers buy more expensive cars of identical quality purely for separation purposes (Bagwell and Bernheim, 1996). At the other end, quality may be over-supplied because it is the focus of comparisons (Bordalo et al., 2016). Intuitively, absent a salient age identifier, consumers would compete for status-rank on car make alone. If premium makes are not in fact higher quality, then the spending is wasteful. However, at the extensive margin, the higher entry price to the status market would deter many marginal status-seekers who would then spend that portion of their budget on a non-conspicuous good. In contrast, if a salient year identifier re-routes the desire to consume conspicuously towards buying newer cars, which on average *are* higher quality than second-hand premium ones - lower emissions, safer for passengers and pedestrians, less likely to break down - then the age identifier could be welfare-improving relative to the counterfactual.

Implicit throughout this paper is the idea that preferences are constructed in response to the environment (Lichtenstein and Slovic, 2006). The decision weights that different consumers place on attributes could be tested through an online experiment that recruits people from different countries, where salience of year information differs. If the interpretation of the results that this paper makes is correct, Irish buyers should place more weight on age (relative to other attributes - make, size, fuel efficiency, mileage) than buyers from other countries.

Insight into the level of salience that drives choices over attributes in the car market could be gained from experiments that manipulate whether the payoff from an

inference depends on coordination or not (i.e. in the spirit of Mehta et al. (1994)). To test primary salience, one could show two cars on screen for a very short period of time, and then ask participants to choose the one whose owner has higher income. To reduce experimenter demand, the target trait might not always be status related (e.g. choose the car owner whose owner is younger/is female/is more environmentally conscious etc.). To measure how attributes perform on secondary salience, participants could be asked to choose the option they thought was the most popular choice by 100 others. Such a test might shed light on how attributes perform as social norms.

To conclude, car purchase is the second largest financial transaction for most households. In most countries the distribution of car sales is nearly uniform, because the factors that trigger purchase are random. The extreme seasonality in Ireland, and to a lesser extent in Great Britain, is difficult to rationalise without giving primacy to the age identifier as a status signal. That status-signalers might place a premium on simple, obvious signals is not accounted for in Spence's signaling framework, where neither attention nor incentives to make correct inferences are limited. But status signalers often want to impress people who are easily distracted, cognitive misers. Behind the figures are thousands of consumers who delayed purchase until a new plate came out, and others who cut short their search to maximise the time period for which they had the 'newest' car on the road. This behaviour might seem frivolous, but it contains important lessons for policymakers considering informational nudges on labels. Government has an unassailable advantage in creating simple, salient labels. This position can be used to minimise externalities and behavioural biases that harm welfare (Bernheim, 2009). Understanding when and why simple, salient labels influence purchase decisions can help policy makers nudge consumers towards patterns of consumption that generate positive externalities. For example, the licence plate might include the carbon emissions per kilometre, or some other numeric measure of "greenness" which lends itself to easy comparison. The evidence in this chpater suggests any green identifier might be more effective if it replaced the age identifier, rather than becoming an additional attribute that competes for limited attention.

4.6 Supplementary Material

4.6.1 Additional Results

Inequality of Sales Pattern in Britain Pre-99

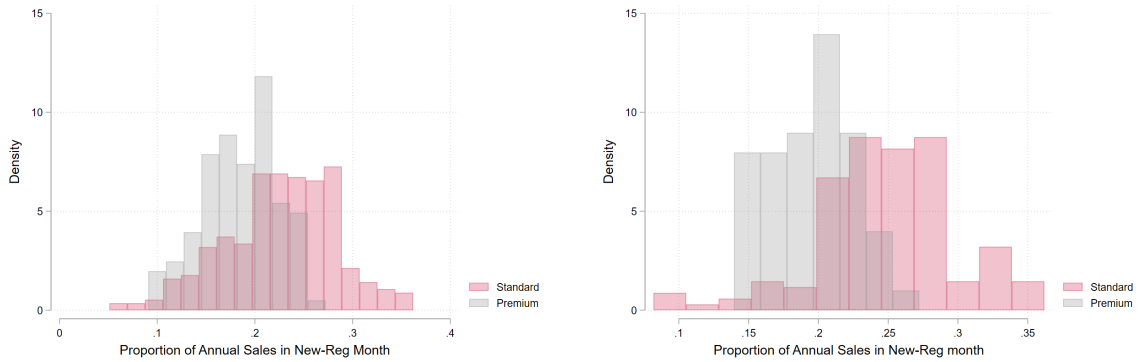

Figure 4.8: Proportion of British sales in New-Reg month (August) when annual plate had salient prefix letter-age identifier. Right panel is only 1990-1998.

Figure 4.8 above shows that pre-99, when the British plate was released annually and the plate began with a letter to indicate the twelve month period (e.g. R for August 1997 to July 1998), a higher proportion of Standard Make sales occurred in the New-Reg month. The assumption that allows this analysis is that cars bought in August will not differ in their survival rate from other months. This allows us to take the distribution of registration dates for each car make in each year as representative.

Ireland: Sales and Import Data

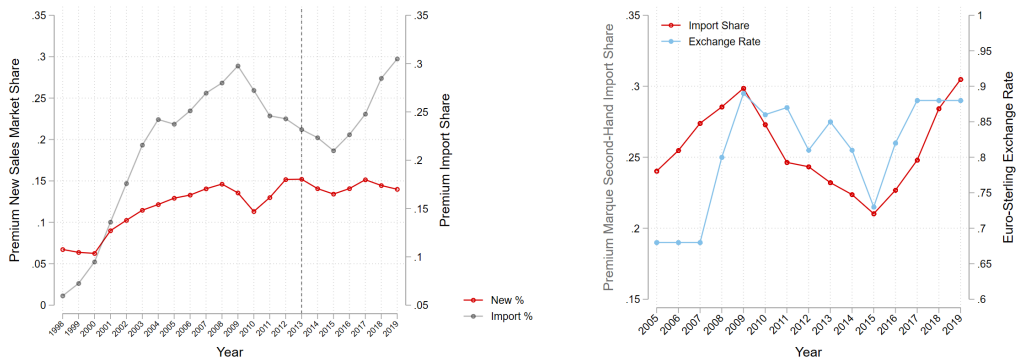

Figure 4.9: Left panel: Premium sales proportion of all new registrations (in red) and share of imports (in grey). Note in 2013 when the treatment occurs, the import share is also falling. By a process of elimination, this suggests demand shifted to the standard-make category. Right panel: The premium-make share of imports against the exchange rate. Note that in 2013, the exchange rate improved but the premium-share still fell (an exception).

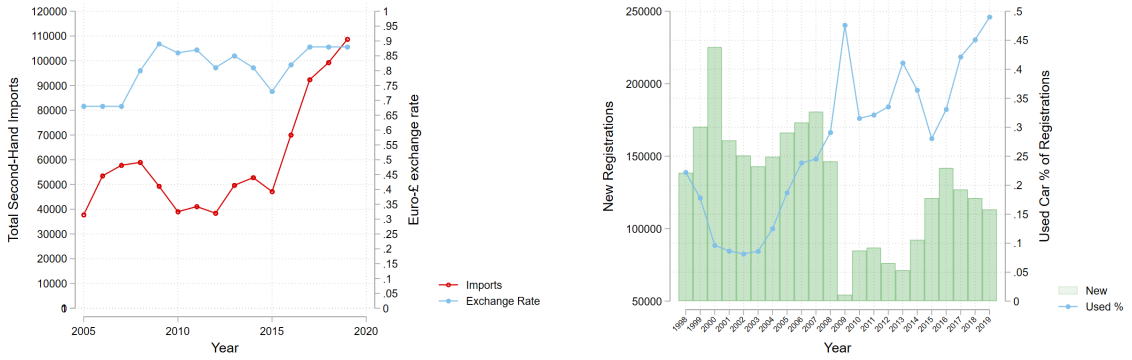

Figure 4.10: Left panel: total number of imports (red) plotted against exchange rate (blue). Right panel: Green bars indicate the number of new registrations each year (note the enormous spike in 2000 and the plummet in 2009 during the Great Recession). The blue line plots imports as a proportion of all registrations.

Additional Regression Results

Tobit Regressions for Make Sales in New-Reg months

	All	Over 0.5 MS	≥ 1% MS	≥ 2% MS
Ireland	0.0730***	0.0740***	0.0761***	0.0828***
	(0.00305)	(0.00285)	(0.00292)	(0.00325)
Premium make	0.0240***	0.0294*	0.00596	0.0155***
	(0.00820)	(0.0156)	(0.00653)	(0.00486)
Ireland × Premium make	-0.0219***	-0.0195***	-0.0190***	-0.0204***
	(0.00455)	(0.00459)	(0.00482)	(0.00550)
Constant	0.144***	0.145***	0.143***	0.139***
	(0.00468)	(0.00434)	(0.00454)	(0.00492)
Observations	1946	1602	1392	1014

Standard errors in parentheses

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

Table 4.7: Tobit Regressions for proportion of make sales in New-Reg months

Northern Ireland Missing Data

Years	Missing Months
2009, 2011, 2012	July, August, September
2014	All
2015, 2016, 2017	October, November, December

Table 4.8: Missing Months of Sales Data for Northern Ireland

4.6.2 Theory

Status Utility and Depreciation

Brand new Premium makes have two sources of status: the make itself (the badge appeal) and the age identifier that comes with it. Total Status utility is a combination of these two attributes:

$$U_s(Age, make) = Age^{\alpha} make^{1-\alpha} \qquad (2)$$

the parameter α in (2) denotes the relative weight placed on the age component of status. Salience theory suggests that α is increasing in the attention paid to this component, which may be determined by presentation factors (the plate format) or be driven by characteristics in the environment (e.g. knowledge level about cars).

The status value of both attributes depreciates. As stated in the main text, the age attribute depreciates faster. Because this is the only source of status for buyers of Standard makes, it implies a higher opportunity cost of delaying purchase within a licence plate period.

Prediction: Purchasing early will be more prevalent for standard brands when the Age attribute is highly salient.

Also, as an aside, note that for the prospective premium buyer who is *uncertain about which car* will allow 'separation' from low types, because she does not know the price range of every car make, there is an incentive to wait and see what others purchase. The average price of a make can be reasonably inferred to be negatively correlated to its frequency (especially when there is a subset of buyers who want to obtain the brand new age identifier at minimum cost) Note that this second-mover strategy is only feasible when age is obvious, so one can focus on new cars when doing the frequency approximation.

Shortening Licence Plate Period

When the licence plate period shortens, a change in the depreciation rate of the Age attribute can drag down the value of make, making the purchase price less appealing given the present-discounted lifetime of status utility one expects to accrue. For instance, consider a person who intends to buy a Premium make and keep it for N years. Let the depreciation parameter take the function form:

$$\delta_i = \frac{1}{rank_i}$$

Under annual licence plates, in the first year, its rank on the Age dimension is 1 (i.e. newest car possible), so there is no depreciation. In the second period, $\delta_i = 1/2$, then $1/3$, and so on. Summing over the intended ownership of the car:

$$\sum_{i=1}^{N} \delta_i Age^{\alpha} make^{1-\alpha}$$

If the licence plate period halves, it creates more periods, which means for a given amount of time there is more depreciation on the overall status value of the car. Assuming the price of the car does not change, this renders purchase of a Premium make a less attractive proposition.

Prediction: Shorter licence plate periods will decrease the market share of Premium makes.

Chapter 5 The Rule of Tome? Longer Novels are more likely to win Literary Awards

Abstract

Longer novels are significantly more likely to win literary awards. This strong and simple bias is shown using all shortlisted novels for three prestigious prizes - the Booker Prize, the Pulitzer Prize for Fiction, and the National Book Award for Fiction, covering a time span of 1963-2021. The result is robust to controlling for author gender and Goodreads rating, and to whether one uses absolute page length or relative length on the shortlist. The size of the effect suggests other valid cues are underweighted in the process of selecting a winner. Judgment and decision making research suggests several causes of the bias. One is the representativeness heuristic: longer novels resemble the tomes that constitute the foundations of the Western canon, and this similarity may subconsciously sway judges. Other explanations include an effort heuristic and the effects of accountability in decisions. These results may explain previous findings that Booker Prize winners are not higher quality than shortlisted novels. The findings cast doubt on the validity of awards as signals of literary merit, and have broader implications for the inferred quality of expert judgment.

This chapter is single-authored. I thank my supervisors, and Deirdre Robertson, Shane Timmons, John O'Hagan and Conor Brennan for helpful comments. This journal version of this chapter is currently revise-and-resubmit at the *Journal of Cultural Economics*.

5.1 Introduction

"Quantity has a quality all its own" - Anonymous

Human organizations and systems rely on expert judgment. In many organizations a tacit belief persists that assembling a group of experts whose incentives are aligned is sufficient to find reasonable answers to difficult questions. The idea is that they will organize the information in a sensible way, give each attribute its normative weight, and wrangle with the trade-offs between options to arrive at a high-quality decision. In popular culture, the best or 'greatest' artworks are chosen each year. This task - imposing an objective ranking on entities whose value is inherently subjective - is a difficult one. Given the task difficulty, coupled with the consequential nature of the decision, most cultural institutions look to experts to decide who should win these awards, rather than deferring to the wisdom of the crowd.

Many artists and interested observers are uneasy about the cultural prominence afforded to awards. One common objection is that awards miss the point of art.[1] A more recent one is that awards rarely go to the highest quality entry, as judged by posterity (Ginsburgh and Weyers, 2014). Part of the unease probably stems from, or is amplified by, the fact that prestigious awards have considerable economic consequences (English, 2014). In literature, winners experience large boosts in sales (Ponzo and Scoppa, 2015; Ashworth et al., 2010). Moreover, lucrative adaptation for film is more likely for winning novels. For instance, the Oscar-winning films *Schindler's List*, *The English Patient* and *Life of Pi* all started as Booker Prize-winning novels. Perhaps more importantly, awards elevate the status of the winner (Frey and Gallus, 2017). Economic benefit may accrue indirectly through this channel. These high stakes motivate quantitative investigation of what matters to the judges.

In this paper I analyze decisions for three English language literary awards: The

[1] Purists might label awards as absurd and pointless and very much like the ribbons to be put away, but this seems unlikely given they serve useful functions, and also, have been around since the start - in the 5th century BC Sopholces and company *competed* in annual theatre competitions.

Booker Prize, the Pulitzer Prize for Fiction, and the National Book Award for Fiction. I present evidence that, conditional on making the shortlist, longer novels are significantly more likely to win these awards. This relationship holds whether absolute or relative length is used. It is also robust to controlling for author gender and Goodreads rating, factors that are plausibly correlated with both novel length and probability of success.

It is reasonable to expect there should be *some* positive association between novel length and success. Writing well is a rare talent; demonstrating that skill over say, four hundred pages or more takes a great deal of effort, and all else equal in the literary stakes, this may tilt the odds in that novel's favor. Length is also positively correlated with less tangible qualities, such as depth, plot complexity, character development. However, I argue that the effect size is *too large* to reflect application of internally consistent preferences. A one standard deviation increase in novel length has a greater effect on success probability than a one standard deviation increase in Goodreads rating. The effect size is larger again when relative length on a shortlist is used. These findings likely constitute evidence of biased decision making by the panel of experts[2] who choose the winner.

Research on decision processes under uncertainty suggests several explanations for the bias. One is that expert judges are engaging the representativeness heuristic. In other words, when faced with the daunting question "which of these novels is greatest?", judges may at least partially, and perhaps subconsciously, be swayed by the answer to a simpler one: "which one most closely resembles what I consider to be a great novel?" The canon of Western literature has novels of formidable length - such as War and Peace, Ulysses, Great Expectations, and Moby Dick - at its foundations. When literary experts are prompted to judge contemporary greatness, these classics may exert some influence. That judges favor longer novels is also consistent with research on magnitude sensitivity. Hsee et al. (2005) found that valuations display

[2]This paper uses the definition of 'expert' suggested by Shanteau (1992) - "those who have been recognized within their profession as having the necessary skills and abilities".

greatest magnitude sensitivity when options are evaluated simultaneously (i.e. joint-evaluation), the evaluability of stimuli is high, and individuals are primed to rely on calculation rather than feeling. These conditions are all met when judging panels compare nominees and choose a winner.

The effect of these biases may be amplified by the group setting. Groups typically outperform individuals in decision-making (see Tindale and Kluwe (2015) for a review) but the group advantage can be diminished by social dynamics within the decision process. In groups comprised of two people, discussion can have a negative effect on accuracy when each does not first produce an independent estimate (Minson et al., 2018). Another potential problem with group decision-making is that when judges have similar training or experience, they tend to share bias and have correlated error (Soll, 1999), a problem that is usually hard to detect, at least from the inside. Accountability and having to justify one's choice to a group can amplify bias in judgments (Lerner and Tetlock, 1999; Slovic, 1975). These channels of potential causation are discussed in Section 5.4.

This paper makes three contributions. The primary one is extending the evidence base for biased expert judgment to wholly subjective domains. This is important because expert judgment is often used as a normative benchmark even when its quality cannot be assessed. For instance, Mollick and Nanda (2016) compared the theater funding decisions of crowds (crowdfunding) to the decisions of experts. The main finding, that crowd decisions are "more wise than mad" was inferred from high levels of congruence with experts' judgments. Without investigating bias in expert judgment, a counterfactual divergence in opinion between crowd and experts might incorrectly be blamed on the crowd. As technologies to aggregate opinions proliferate, it is vital that the measure against which these new methods are compared is itself valid. Existing evidence of deficiency in subjective expert judgment is sparse.[3] One exception is Ginsburgh and Van Ours (2003) which showed that in a prestigious

[3]The majority of research in expert judgment uses forecasting tasks (or similar) where it is straightforward to measure objective performance (e.g. Tetlock, 2009; Camerer and Johnson, 1991).

music competition, randomly-assigned order of performance had a significant effect on judges' ranking, and hence subsequent commercial success. Those who performed later in the competition were ranked more highly. The authors noted that this finding "sheds some doubt on their [the judges] ability to cast fully objective judgments." (Ginsburgh and Van Ours, 2003, p. 294)

The second contribution is providing a possible explanation for the finding of Ginsburgh (2003) that literary awards do not select aesthetically superior works. Ginsburgh composed a measure of the long-term reputation of the novels, assumed to capture fundamental aesthetic quality. He showed that prizes are often poor predictors of true aesthetic quality or survival of the work. However, no explanation was offered for this poor predictive power. The results of this paper potentially fill that gap. Specifically, a wedge will be driven between what judging panels favour and what stands the test of time if evaluation of *relative quality* is biased by factors that do not afflict individual judgment of individual novels.

Third, this paper extends the nascent research on the downstream effects of the unconscious biases of literary gatekeepers. Awarding institutions are creators of cultural value (English, 2014). Bias in the judging process distorts what is considered culturally valuable. Existing evidence of bias comes from analysis of New Yorker short stories, where editors favored short stories that featured protagonists with whom they shared traits (Milkman et al., 2007). This preference for familiarity perpetuates exclusion of historically underrepresented perspectives. Bias in literary award choice propagates through a commercial channel: Under the counterfactual of unbiased judgment, the public would be encouraged, on average, to read shorter novels of greater literary merit. The bias also works through a professional status channel. Correlational evidence suggests the decisions of the literary judging panels are taken to be sound by their peers, acting as a diagnostic cue for who deserves the highest award for literary merit - the Nobel Prize for Literature. Since 1975, of the nine English language novelists who have won the Nobel, seven had previously won either

the Booker, Pulitzer or National Book Award.[4]

5.2 Data

This section first summarizes the history of three literary awards that are analyzed, and then lists the details of the shortlisted novels that were gathered during data collection.

5.2.1 Literary Awards

Since its inception in 1969, the Booker Prize has been awarded each year to the best original full-length novel written in English. Until 2014, only writers from the Commonwealth of Nations or the Republic of Ireland were eligible. In 2014, eligibility was expanded to any work published in the United Kingdom and written in English. The largest cohort of previously excluded writers were U.S novelists. In its 54-year history, 315 novels have been shortlisted – usually six per year - and 58 have won. The convention of a single winner was broken in 1970, 1974, 1992 and most recently in 2019, when two novels were declared joint winners.

The Pulitzer Prize for Fiction, which was inaugurated in 1917, recognizes "distinguished fiction by an American author, preferably dealing with American life", published during the preceding calendar year. The Prize began naming the shortlist in 1980. The shortlist usually consists of three novels, but on rare occasions two or four. It is unlike the Booker in three respects. First, the award has never been shared. Second, on eleven occasions no prize was awarded. And unlike the Booker, it has

[4]One exception is Doris Lessing (Nobel winner in 2007), who was nominated three times for the Booker Prize. Only once did she lose to a shorter novel: in 1971 her entry was 278 pages long and the winner's was 247 (that winner, V.S Naipaul, also went on to win the Nobel). The other exception is the 2021 winner, Abdulrazak Gurnah, who was nominated for the Booker in 1994 for his novel *Paradise* (length: 256 pages), which lost out to *How Late It Was, How Late* (length: 384 pages). See Table 5.4 in Supplementary Material for a full list.

maintained a strict nationality rule.

The National Book Award for Fiction (NBAF) is one of five annual National Book Awards, which recognize outstanding literary work by United States citizens. The NBAF began in 1950, with a shortlist of ten novels. The shortlist was changed to six novels in 1963, which is the point the current analysis begins. Shortlist length fluctuated in the following decades, but since 1990 it has always been made up of five novels. The website of the National Book Awards note that they are awards "by writers to writers" and that the judging panel are five "writers who are known to be doing great work in their genre or field" (National-Book-Awards, 2021).

5.2.2 Data Collection

Information about each shortlisted novel was taken from Goodreads.com. The latest awards year is 2021. The final dataset contained the following variables: year, author, title, novel length in pages ('length')[5], outcome (1 = winner, 0 = loser), Goodreads mean rating, and author gender. Nationality was also recorded for the Booker prize shortlists, as this was the only international award.

5.3 Results

5.3.1 Summary Statistics

Figure 5.1 below shows a scatter plot of page length for winning novels and losing nominees against time. A density plot of page length is shown in the right panel. The mean length of winners is 387 pages (SD = 163) and the mean length of losing novels is 332 pages (SD = 143). This difference is highly significant whether a parametric

[5]Would the results would be different if word count was used instead of number of pages as the length variable? It seems reasonable to assume that longer books are more likely to use smaller font, to reduce production costs. A negative correlation between word count and font size would imply the effect size using the number-of-pages length proxy is a lower bound.

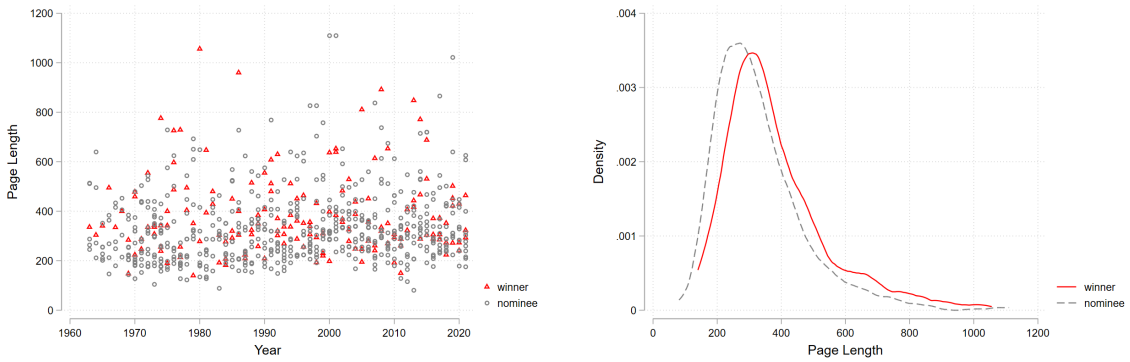

Figure 5.1: Left panel: Scatter plot of year against page length for winners (red triangles) and losing nominees (grey circles). Right panel: Density plot of page length for winners (in red) and losing nominees (grey dashed line).

test (two-sample t-test, t = -4.22, df = 750, $p < 0.0001$, Cohen's d = 0.372) or non-parametric test is used (Wilcoxon rank-sum test, z = -4.46, $p < 0.0001$).

Summary statistics for the three awards are shown in Table 5.1 below, with the rightmost two columns showing the t-statistics from a parametric test of difference-in-means and the z-score from a non-parametric Wilcoxon rank sum test. Note the smaller sample size for the Pulitzer Prize (due to only three novels on the shortlist each year) means the difference is only marginally significant for this award, despite the point estimate being the same magnitude.

Award	Obs.	Winners	Losing Nominees	Difference	T-test (t)	Rank-sum (z)
Booker	315	370	321	49	2.36**	2.52**
Pulitzer	139	416	358	58	1.84*	2.04**
NB Award	298	379	333	46	2.29**	2.52**

Table 5.1: Summary Statistics for three literary prizes. Winner column denotes mean length of winning novels, loser denotes same for losing novels. Significance levels: * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$.

5.3.2 Logistic Regression

This section analyzes how the probability of winning changes with length using a logistic link function. The results are shown in Table 5.2 below. The first column, Model A, shows the most simple specification, where the standardized length of the novel is the only predictor. The odds ratio of 1.387 on this variable shows that a one standard deviation increase in length increases the odds of winning by almost 40%, and the effect is highly significant ($p < 0.0001$). To interpret this effect, think of a five-novel shortlist where one initially assigns equal chance to all contenders i.e. odds of 4/1 or a 20% chance of success. The 1.38 odds ratio point estimate implies that that 20% chance of success would rise to 28% for a novel that was one standard deviation longer. The model includes dummy variables for award type and fixed effects for year, to control for the variation in shortlist size, which alters the base probability of winning, and also for any possible changes in judges' preferences over time.

	A	B
Length (standardized)	1.387***	1.31***
	(0.126)	(0.123)
Female author (ref. = male)		0.64**
		(0.132)
Goodreads rating (standardized)		1.27**
		(0.139)
Year and Award Fixed effects	Yes	Yes
Observations	744	743
Exponentiated coefficients; Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$		

Table 5.2: Logistic Regression of binary award success variable on novel and author characteristics.

However, the apparent relationship between length and award success could be spurious. An omitted variable positively correlated with length might be what judges actually favor. For example, male writers might be preferred by judges and also just happen to write longer novels. Alternatively, longer novels might simply be longer because they are better e.g. these authors' editors recommended fewer cuts. If this is

the case, the predictive power of the length variable should be greatly reduced once the Goodreads rating is added as an explanatory factor for the outcome. The results of Model B show that the coefficient on length reduces only slightly when gender and rating are added, and the effect is still highly significant (p = 0.004). A bias against female authors is significant (p = 0.029). Table 5.2 shows a point estimate of 0.64 on the odds ratio, which implies female authors being two-thirds as likely to win as male authors, controlling for other observable factors. Lastly, higher Goodreads ratings are associated with improved odds of success. The relationship is significant at the five percent level (p = 0.029). Caution is required interpreting this relationship, as the public might give higher ratings following the opinion of the judges who granted the award (Jacobsen, 2015). In other words, this coefficient might be inflated due to reverse causality (the Goodreads rating is high *because* it won the award).

The significant positive relationship is also present if one uses a linear probability model (LPM) instead of a logistic-link function to model the outcome. In an LPM, the Model A length coefficient implies a six percentage point increase in the likelihood of winning for a one standard deviation increase in length (β_{Length} = 0.0596, se = .016, p < 0.01). In Model B, female authors have a decreased probability of winning of just under seven percent points (β_{Female} = -0.0685, se = .033, p < 0.05). The Goodreads rating has a significant positive association with winning but the effect size is smaller than that of length and gender ($\beta_{Goodreads}$ = 0.034, se = .0174, p <0.05). The full LPM output for Models A and B is shown in Table 5.5 in the Supplementary Material.

5.3.3 Relative Length on Shortlist

A metric to calculate the effect of relative length of novels on each shortlist was constructed by calculating the mean length of each shortlist, then computing the difference between each novel's length and this mean. This was also done for the median length. The resulting distributions, shown in Figure 5.2 below, did not pass the Shapiro-Wilk test for Normality. As a result, each variable was divided into five

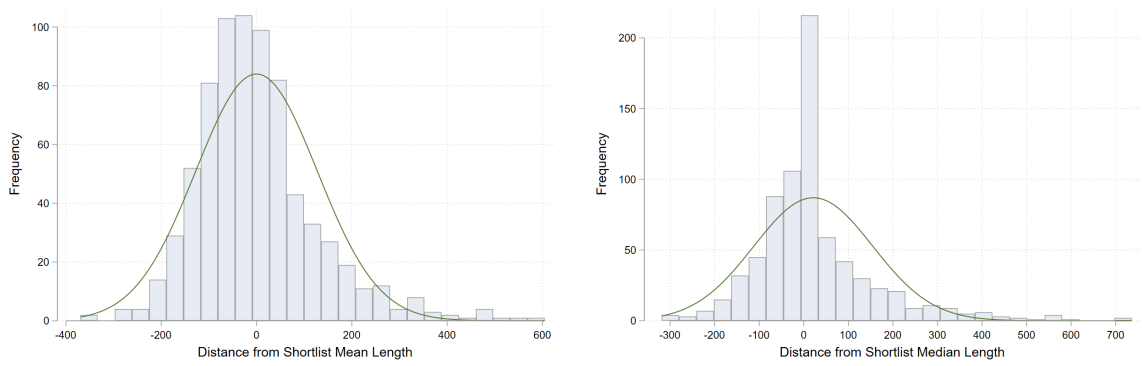

Figure 5.2: Distributions of relative length on shortlist with Normal distribution overlaid. Distance from shortlist mean length (left panel) and distance from shortlist median length (right panel). Note the large spike at zero is due to each shortlist with an odd number of nominees having one novel at exactly the median length (by definition).

quintiles and the award outcome was modeled on this categorical variable.

	Mean	Mean	Median	Median
Quintile 20-40 (ref cat. = Quintile 0-20)	2.34**	2.583**	1.68	1.75
	(0.82)	(0.947)	(0.586)	(0.639)
Quintile 40-60	2.78***	3.12***	2.21**	2.17**
	(0.959)	(1.14)	(0.735)	(0.751)
Quintile 60-80	3.67***	3.99***	4.06***	4.08***
	(1.237)	(1.41)	(1.303)	(1.36)
Quintile 80-100	3.55***	3.25***	2.83***	2.40**
	(1.20)	(1.13)	(0.927)	(0.818)
Female		0.623**		0.644**
		(0.131)		(0.137)
Goodreads Rating (standardized)		1.31**		1.27**
		(0.145)		(0.141)
Year and Award Fixed Effects	Yes	Yes	Yes	Yes
Observations	744	743	744	743

Table 5.3: Logistic Regression: Dependent variable is award success (0 or 1). Quintiles are categorical variables that denote how far each observation is from the average length of that award-year shortlist.

Table 5.3 shows the results. The coefficients are displayed as odds-ratios. Column

1 and 2 show the effect of distance-from-the-mean quintiles on success. Column 1 shows that novels in the longest two quintiles are 3.5 times more likely to win than the novels in the shortest quintile (the reference category). Column 2 shows this relative length effect dwarfs the effect of the Goodreads rating. As before, female authors are just under two-thirds as likely to win ($p < 0.05$). Columns 3 and 4 employ quintiles based on distance from the median length. The pattern of results is more pronounced, with no significant increase in the odds of success for novels in Quintile 20-40, but a larger advantage for novels in the Quintile 60-80. This suggests the judges may use the median length as a reference point - which would make sense, as this statistic is more accessible than the mean.

When the reference category is changed to the middle quintile (40-60), the odds ratio coefficient on the shortest quintile (0-20) in Columns 1-4 is 0.36, 0.32, 0.45, and 0.46 respectively ($p < 0.05$ in all cases). Comparison of point estimates shows that the bias against the shortest novels is apparently larger than the bias against female authors (see Table 5.7 in Supplementary Materials for details).

Length Rank on Shortlist

Another way to measure relative length on the shortlist is to simply use the rank. However, shortlists of different lengths (Booker = 6, Pulitzer = 3, NBA = 5) complicate this slightly. In Table 5.8 in the Supplementary Material, award success is modeled using tercile rank for the combined Booker and Pulitzer sample. The results are in concordance with the evidence presented above. Relative to novels in the longest tercile, middle-tercile novels are half as likely to win ($\beta = 0.536$, $p < 0.05$), and shortest-tercile novels are one-third as likely to win ($\beta = 0.356$, $p < 0.01$). When author gender and Goodreads rating are added, the middle tercile coefficient increases to 0.644 and is no longer significant, but novels in the shortest tercile remain significantly less likely to win ($\beta = 0.433$, $p < 0.01$).

Table 5.8 also regresses success on a continuous length-rank variable for the combined

Booker and National Book Award sample, where the longest novel has a rank of one. As expected by now, an increase in rank is associated with a lower probability of success ($\beta = 0.77$, $p < 0.001$).

Taken together, this shows that whether absolute page length, relative length in pages, or length rank is used, the relationship between success and novel length is robust.

5.4 Potential Causes of the Bias

There are several potential causes of the bias towards longer novels. One is the use of the representativeness heuristic. Part of the Kahneman and Tversky (1972) definition of representativeness is "the degree to which it [the object of appraisal] is similar in essential characteristics to its parent population". Longer novels physically resemble the 'great' novels in the canon, and this similarity may sway judges who are tasked with identifying contemporary greatness in a complex multiattribute choice task. Moreover, somewhat counterintuitively, experts might be more susceptible to this bias than average readers, because they are more familiar with the weighty classics in the canon.

Representativeness may be exacerbated by the nature of the judging context, where members of the group are accountable, combined with the fact that novel length is a physically prominent attribute. Such attributes are given greater weight when choices need to be justified to others (Slovic, 1975). The effect is stronger in groups than for individuals (Irwin and Davis, 1995). It occurs because group members prefer "noncontroversial reasoning" to explain their choices, and an argument to choose the alternative that is highest on a prominent dimension "is both easy to explain and likely to be accepted by a majority of group members" (Irwin and Davis, 1995, p. 329; see also Barber et al., 2003). Accountability increases the desire to avoid appearing foolish, and therefore a preference for options that are easy to justify (Lerner and Tetlock, 1999). The desire to not appear foolish might be particularly strong in the

rarefied air where literary awards are decided. Moreover, Lerner and Tetlock (1999) highlighted that accountability tends to exacerbate the naive use of normatively (but not obviously) irrelevant cues (Tetlock and Boettger, 1989). Of course, *length in itself* is unlikely to be cited as the reason to favor one novel over the others, but - speculatively - euphemistic terms such as 'depth', 'complexity' and 'broad sweeping narrative, beautifully wrought' may be used instead.

That judges favor longer novels is also consistent with research on magnitude sensitivity, sometimes referred to as scope sensitivity. Hsee et al. (2005) found that valuations display greatest magnitude sensitivity when options are evaluated simultaneously (i.e. joint-evaluation), the evaluator feels competent to appraise each option (evaluability is high), and evaluation is primed to rely on calculation rather than feeling. These conditions are all met when judging panels compare nominees and choose a winner. In contrast, under single-evaluation, and when evaluability is low (Morewedge et al., 2009), decision makers are relatively insensitive to magnitude. Sensitivity to magnitude can be harmful to decision quality, as judged by experienced utility: as the authors summarised "[I]n decision making, more often seems better, yet in life, more is often not better" (Hsee et al., 2005, p. 237).

A different mechanism that is also plausible is use of the effort heuristic, wherein the perceived effort of producing a cultural work is used as a proxy for its quality (Kruger et al., 2004). At a fundamental level, this heuristic appears a specific manifestation of the general tendency to overlook subtractive changes that improve quality (Adams et al., 2021). The effort of reading the work may be used as a diagnostic cue too, which may be amplified by the group context. Judges might be concerned that citing a preference for the shortest novel would be misconstrued as laziness by their peers or evidence of not engaging properly in the longer works.

Alternatively, the bias could be produced by the way judges carry out the evaluation process. For instance, judges might choose to read the shortest novel first to 'get off the mark' (this is partly based on introspection). This could give rise to a sequence effect

if initial expectations for quality are too high but then quickly adjust downwards (Criscuolo et al., 2021). As outlined in the introduction, sequence effects have been documented in a prestigious piano award where later slots were more advantageous (Ginsburgh and Van Ours, 2003). Alternatively, different judges may seek evidence to support an early first impression i.e. confirmation bias. If this first impression crystallizes at an absolute point in the novel, perhaps after one hundred pages, and judges spend the rest of the novel seeking confirming evidence, longer novels will be more strongly favored, and perhaps more robustly defended in the discussion to determine the winner. In the jargon of marketing science, judges may apply a (motivated) attribute-adding evaluation strategy, rather than attribute-averaging (Troutman and Shanteau, 1976).

It should be noted that little research has been done on the judging process itself, and consequently the explanations above should be regarded as tentative.

5.5 Conclusion

Important organizational decisions are likely to be made by groups rather than by individuals (Tindale and Winget, 2019). This paper demonstrates a strong and disarmingly simple bias in group judgment. When literary experts come together to dispute tastes, shorter novels consistently get the short shrift. This is surprising, because *conditional on making the shortlist*, one would not expect length to be a decisive factor, in much the same way that conditional on making NBA league, one would not expect player height to predict scoring rate (it doesn't). The significant association is robust to controlling for other factors such as author gender, Goodreads rating, and is not due to judging preferences in a particular decade. The absolute size of the bias against shorter novels is similar in magnitude to the bias against female authors. The gender bias in literary awards has been noted elsewhere (Moseley et al., 2019), but this paper is the first to quantify it (though its purpose here is mainly comparative).

The most likely causes of the bias were discussed in the previous section. Attribute substitution, which is the basis of the representativeness heuristic, is common whenever objective criteria for making a difficult decision are scarce (Kahneman and Frederick, 2002). For this reason representativeness was given precedence. Magnitude sensitivity under joint evaluation in a deliberative or calculating mindset is also plausible, given past findings. Social dynamics - particularly accountability to panel colleagues - may have exacerbated the tendency to favor longer novels on shortlists. For the practical purpose of understanding why Booker Prize winners are not higher quality than nominees (Ginsburgh, 2003), it is not necessary to identify the precise cause of the bias. However, this would be a logical step for future research. The task of literary award panels could be recreated with reasonable external validity with appropriate sampling of participants. Beyond literary awards, the interaction between magnitude sensitivity and reason-based judgments, which is common in a group context, may prove a potentially fertile research area of general interest. For instance, in hiring decisions, is the quantitative measure of years-of-experience given disproportionate weight? In the sports transfer market, do metrics like career goals/points scored receive more weight than qualitative cues, such as the opinion of a coach or scout? The advantage of these domains is that feedback on objective decision quality could be obtained.

What should the awarding bodies make of these findings? Providing a signal of quality is a practical and important function of literary awards, because novels are experience goods, whose quality is impossible to ascertain prior to consumption. The integrity of awards depends on this signal being valid. One option would be to incorporate best practices from decision analysis to improve the quality of judges' decisions. For instance, alerting decision makers about the risk of attribute substitution can prevent judgment being swayed by a marginally relevant variable (Schwarz and Clore, 1983). A 'literary merit' checklist or scorecard might also be beneficial in reducing the chances that prominent attributes receive more than their normative decision weight (Milkman et al., 2015). Of course there would still be scope to debate how to

combine the various attributes to form an overall evaluation. Altering the decision environment is another way to improve decision quality (Klayman and Brown, 1993). To maximize the value of collaboration, independent judgments need to be recorded before discussion takes place. Otherwise peer opinions exert a gravitational pull that diminishes the advantage of two heads over one (Minson et al., 2018).

Alternatively, awarding bodies may opt for more radical changes. The Academy Awards uses a wisdom-of-the-expert-crowd selection process, in which just under 10,000 members vote anonymously. This process is immune to biases produced by the group context of the judging panel (e.g. accountability), although it has its own pitfalls. The time investment to read a novel is so much greater than to watch a film that this system might not seem feasible for literary awards. However, with enough judges, it would not be necessary for each to read every novel.[6] Lastly, if the implementation costs of these suggestions seems to outweigh the benefit, there is always the simpler option presented by Ginsburgh and Weyers (2014): announce a list of winners, rather than a single one.

To conclude, the highlighted bias should inform evaluations about the quality of expert judgments in literature. It may also prompt reflection on how norms in cultural production emerge and are sustained. The broader relevance of this research is that expert panels are entrusted with making important decisions across a range of domains. The current findings should stimulate investigations into how social dynamics influence the decision weight placed on quantitative attributes in complex subjective choice tasks.

[6]Consider if a large number of judges were randomly allocated a subset of the nominations and submitted scores. This seems just as reasonable as the current setting, where five judges create a shortlist from approximately 150 contenders before deciding on a winner.

5.6 Supplementary Material

Nobel Prize

As stated in the introduction of the main manuscript, seven of the nine English language novelist winners of the Nobel Prize has previously won one of the three awards analysed in this paper. The full list is presented in Table 5.4 below. Correlation is not causation, but the pattern is suggestive that winning a literary prize makes one appear 'eligible' for the Nobel.

Winner	Literary Award	Year of Nobel Win
Saul Bellow	NBAF (1971, 1976) and Pulitzer (1976)	1976
William Golding	Booker (1980)	1983
Nadine Gordimer	Booker (1974)	1991
Toni Morrison	Pulitzer (1988)	1993
V.S Naipaul	Booker (1971)	2001
J.M. Coetzee	Booker (1983, 1999)	2003
Doris Lessing	Nominated for Booker in 1971,1981, 1985 but never won	2007
Kazuo Ishiguro	Booker (1989)	2017
Abdulrazak Gurnah	Nominated for Booker (1994)	2021

Table 5.4: List of English-language Nobel Prize for Literature winners since 1975. Note 7 of 9 also won one of Booker, Pulitzer or National Book Award prior to Nobel win.

Linear Probability Model

	A	B
Length (standardized)	0.0597***	0.0498***
	(0.0165)	(0.0169)
Female author		-0.0685**
		(0.0329)
Goodreads rating (standardized)		0.0343**
		(0.0171)
Constant	0.147	0.167
	(0.175)	(0.174)
Year and Award FEs	Yes	Yes
Observations	744	743
Adjusted R^2	-0.048	-0.037
Standard errors in parentheses		
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$		

Table 5.5: Linear Probability Model (discussed in Section 5.3)

Relative Length with different reference category

Table 5.6 below presents the relative length analysis (Section 5.3.3 of main text) with the middle quintile as the reference category (Quintile 40-60). The effect of relative length is increasing but not monotonic.

	Mean	Mean	Median	Median
Quintile 0-20	0.360***	0.321***	0.452**	0.438**
	(0.124)	(0.117)	(0.151)	(0.150)
Quintile 20-40	0.842	0.828	0.761	0.785
	(0.243)	(0.253)	(0.231)	(0.247)
Quintile 40-60 (ref cat.)				
Quintile 60-80	1.321	1.279	1.838**	1.827**
	(0.358)	(0.363)	(0.499)	(0.527)
Quintile 80-100	1.279	1.043	1.279	1.057
	(0.349)	(0.305)	(0.357)	(0.308)
Female author		0.623**		0.618**
		(0.131)		(0.129)
Goodreads rating (standardized)		1.314**		1.325**
		(0.146)		(0.148)
Year and Award FEs	Yes	Yes	Yes	Yes
Observations	744	743	744	743

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5.6: Relative length regression with middle quintile as reference category.

The main manuscript said that relative length bias was stronger than the bias against female authors. This can be seen by comparing the coefficients of Quintile 0-20 and Female author in Columns 2 and 4 above. However, the statement needs to be qualified. The female bias can only be measured to the extent that female authors make the shortlist and do not win. But across all 150 award-year shortlists, fifteen contained no female authors. Hence the estimate of this bias may be a lower bound. Male authors were a majority on almost three quarters of shortlists (73%). Table 5.7 shows the full breakdown.

Number	Freq.	Percent	Cum.
0	15	10.00	10.00
1	50	33.33	43.33
2	44	29.33	72.67
3	26	17.33	90.00
4	13	8.67	98.67
5	1	0.67	99.33
8	1	0.67	100.00

Table 5.7: Number of female authors on shortlists. Note that in 1970, there were 12 novels shortlisted for the Booker Prize, of which eight were by female authors.

Shortlist length rank

As discussed in section 5.3.3 of the main text, using length rank is a simple way to measure the effect of relative length on success. However, shortlists of different lengths (Booker = 6, Pulitzer = 3, National Book Award = 5) makes this approach slightly awkward. In Table 5.8 below, tercile rank of length is used as the explanatory variable for the combined Booker Prize and Pulitzer sample (categorizing the National Book Award shortlist into terciles would involve making arbitrary assumptions). To be explicit, the assignment works as follows: The two longest books on the Booker shortlist are in the longest tercile, the third and fourth longest are in the middle tercile, and the shortest two are in the shortest tercile. The Pulitzer has a shortlist of length three so rank automatically maps to tercile. As discussed in the main text, Columns 1 and 2 in Table 5.8 show that the middle and shortest terciles have lower odds of success. The Column 2 specification adds author gender and Goodreads rating. The point estimate for the female author e variable is not significant, but Goodreads rating is a positive predictor of success ($p < 0.05$).

	(1)	(2)	(3)	(4)
Middle tercile (ref cat. = longest)	0.536**	0.644		
	(0.150)	(0.187)		
Shortest tercile	0.356***	0.433***		
	(0.109)	(0.138)		
Length rank (continuous)			0.770***	0.781***
			(0.0474)	(0.0490)
Female author		0.676		0.592**
		(0.181)		(0.143)
Goodreads rating (standardized)		1.423**		1.123
		(0.227)		(0.132)
Year and Award FEs	Yes	Yes	Yes	Yes
Observations	444	444	613	612

Exponentiated coefficients; Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5.8: Effect of rank on probability of success.

In Columns 3 and 4 length rank is included as a continuous variable for the Booker and National Book Award sample. The result is similar to other specifications, with the probability of success falling as rank increases. The effect is highly significant ($p < 0.01$). In a reversal of the Column 2 pattern, in Column 4 the female author variable is significant ($p < 0.05$) but the Goodreads rating is not. Taking these results in conjunction with the other model specifications, it seems reasonable to conclude that novel length is the variable with the most stable predictive power for award success.

Chapter 6 Conclusion

The introduction to this thesis invoked the call in Simon (1983) to interrogate the taken-for-granted assumptions of economic models. One such assumption relates to how numerical information is perceived and integrated into decisions. Concomitant with the concept of the rational agent is the assumption that numerical information is represented accurately, i.e. that perception of number matches the objective reality. However, research in numerical cognition demonstrates that perception is distorted, with subjective perceptions compressed as stimulus magnitude increases. This is a common feature of sensory modalities. Should we care? This thesis set out to investigate the practical implications of the "number sense" for economic decision making. The first section of this conclusion summarises the insights from the thesis on this core research question. I also discuss limitations in terms of method and data. The second subsection discusses avenues of interest for future research stemming from the findings on underestimation bias (Chapter 2 and 3), cognitive limitations and energy efficiency (Chapter 3), status utility and salience (Chapter 4), and the quality of decisions made by groups (Chapter 5).

6.1 The "Number Sense" and Decision Making

At the outset of this thesis, the state-of-knowledge regarding the accuracy of intuitive summation was as follows: in unincentivised tasks, people tended to underestimate the value of their shopping basket, and in one decision-from-experience task students underestimated (on average) the sums of rapidly-presented number sequences (Scheibehenne, 2019). It was not obvious that this bias would remain when both incentivisation and experimental control over numeric stimuli were imposed. Nor was it obvious, or even probable, that the bias would manifest in a forced-choice task. But the results in Chapter 2 showed the effect did generalise. The effect size of approximately 6% underestimation was in line with previous work, which is evi-

dence in favour of it being a robust phenomenon. The inference we can make about the existence of the bias is quite strong due to the experimental method deployed. Ascertaining the external validity of the bias is a different issue, which is taken up in Section 6.2.

Chapter 3 described a valuation task designed to identify whether underestimation bias could underlie choice behaviour generally attributed to time preferences. A procedural difference from Chapter 2 was that a sequence of numbers was not presented to participants. Instead, they were given a difference for fuel costs, and we hypothesised that they would underestimate what this difference amounted to over the given time period. This expansion of the circumstances in which the bias applies is justifiable if the compressive mental number line affects estimates of sums generated by intuitive addition *or* by multiplication. The Chapter 3 experiment also tested concentration bias, a similar but distinct bias that stems from the limited-attention dimension of bounded rationality. The pattern of results suggested that both biases can play a role in energy-efficiency investment decisions. However, compared to Chapter 2, caution is required in drawing inferences from the results of Chapter 3. The task was hypothetical in nature, and moreover time constraints meant it was not possible to include attention or comprehension checks. Some participants may have misunderstood the task. These aspects probably explain the high level of variation in WTP responses. The separating test between the mechanisms was also inconclusive. Had more time been available for the experiment, an estimation task (similar to those in Chapter 2) could have been conducted, and the level of underestimation at the individual level used as a predictor variable for the WTP responses. Despite these design limitations, the results are useful for informing future research on how to diminish the energy-efficiency gap.

Chapter 4 also provided an indirect test of a core property of the number sense, namely that the perceived change in numerosity is proportional to the initial stimulus. In other words, an increase from 1 to 2 is subjectively greater than from 98 to 99. A unique feature of the Irish licence plate system is that the registration begins with

the last two digits of the year. The prediction from numeric cognition is that the step-change in 'newness' between an 02 and 01 registration will be greater than between a 99 and 98. The difference between 99 and 00 should appear greatest of all. The empirical results appear to be broadly consistent with this prediction. Starting in 2000 there was a marked increase in the proportion of annual sales that occurred in January. The implication is that lower numbers made the subjective gap in 'newness' appear greater, and hence increased the appeal of early purchase. Of course, Simon's warning applies here too, and one must always consider alternatives. An explanation for the pattern that requires no distorted "number sense" is that rising prosperity during the Celtic Tiger years increased the weight given to social status in the utility function. This would have enhanced the payoff from conspicuous consumption and led to an increase in the proportion of annual sales that occurred soon after the new plate was released. Or, an initial shift in consumer behaviour may have spurred a supply-side response (e.g. increased advertising of special offers in January) which may have amplified the demand-side shift. The existing data is insufficient to identify the primary cause of the increased peak in seasonality of demand.

Chapter 5 did not test the "number sense" directly, but the broader concept of the decision weight given to a quantitative variable, and hence its findings are not directly relevant to this subsection. Several explanations were forwarded for the observed positive correlation between book length and award success. One of the candidate mechanisms, namely accountability, suggests the group context may increase the weight on quantitative information, even if this higher weighting is tacit. But richer data, and perhaps even complementary experimental data mimicking the task, are necessary to isolate the primary channels of causation.

6.2 Future Research

This section discusses directions for future research. Some are designed to address limitations of the current research; others are research questions of interest that naturally arise from the the current findings.

Underestimation Bias

To reiterate, the results of Chapter 2 suggest underestimation bias may generalise across a range of consumer choice domains. The bias may have been hiding in plain sight for some time. Within the domain of estimating utility bills, a sensible next step may be to investigate the process by which people decide how to direct savings efforts. It seems plausible that people have an approximate internal ranking of their annual expenditures. If intuitive estimates of sequence sums undershoot their true totals by even 5-10%, recurring expenses might be downgraded in this internal category ranking. This could mean switching efforts are directed elsewhere, to areas where potential savings are smaller, implying economic loss. Direct experimental or field evidence would be useful in clarifying how people keep track of expenditures and choose which markets to expend search-and-switch costs in the hope of a worthwhile return. Evidence of this nature could inform the policy response to design and proliferate useful decision tracking tools to help boost decision makers' capabilities (Hertwig, 2017).

When consumers *pay more over time* for a product or service that could be bought upfront, the usual explanation is liquidity constraints. But underestimation bias points towards payments over time being perceived as less costly than they truly are. If the perceived price is lower, demand may be shifted upwards, with the implication that the bias can be a profitable one for companies to exploit. Some Fintech companies act as intermediaries for firms who want to sell products in instalments. Those who select into these services are plausibly at higher risk of underestimation bias. If that is the case, provision of these services may warrant closer attention from consumer

protection agencies.

Relatedly, underestimation bias may also contribute to explaining consumption patterns currently attributed to overly-optimistic beliefs about the future. For instance, in a highly influential paper titled "Paying not to go to the gym", DellaVigna and Malmendier (2006) showed that gym members who pay a flat-rate monthly membership would save money by buying 10-visit passes instead. The difference in ex-post price-per-visit was attributed to excessive optimism about likely usage by monthly members. However, underestimating the total cost of annual membership at the initial point of purchase (when the monthly fee is described) could also contribute to the effect. In the sample of gyms used, the monthly-payment contract cost more per year than the annual payment option (e.g. $85 per month vs. $850 per year). Anyone who underestimated the annualised cost of the monthly payment option would pay more per-visit than they expected to, even when their beliefs over attendance frequency were correct. This could contribute to the apparent puzzle of why 10-visit passes for $100 were not more popular.

Cognitive Limitations and Energy Efficiency

The theoretical motivation for the experiment in Chapter 3 was the lack of attention given to possible cognitive components of the energy-efficiency gap. The WTP experiment tested the predictive power of two cognitive mechanisms, underestimation bias and concentration bias. The separation test between these biases was inconclusive. Future experiments could implement better designs to test their relative predictive power, using a within-subjects design and collection of important controls such as individual discount rates. It should also be noted that at the time of designing the experiment, concentration bias was the leading model of how limited attention could distort choices in energy-efficiency investments. It has since emerged that relative thinking (Bushong et al., 2021) and salience models (Bordalo et al., 2021) have better predictive power in consumer choice settings (Somerville, 2022). These alternative

explanations could be tested in future.

From an applied point of view, the motivation for the experiment in Chapter 3 was that the climate crisis requires drastic action. The main result of interest to policy makers is that WTP for energy-efficiency in a leasing scenario is higher when the payments are spread out over the 36 months of the contract. Interestingly, one solution in Hausman (1979) to the high implied discount rates was that utility companies should purchase efficient technologies and lease them to households: "Presumably utilities would be willing to engage in such profitable activity, since they could borrow money to finance the more energy-efficient appliances and then charge a rental rate which would leave the consumer better off". Hausman dismissed the idea in the same paragraph: "a clear incentive problem exists because more efficient models might lead to a decreased electricity demand, which is the primary product of those same companies" (pg 52). But cannibalisation of revenue streams does not apply to car companies, who typically do not sell fuel too. They could lease fuel-efficient cars to consumers and charge a rental price that leaves both better off. The difference in WTP by payment schedule was particularly large for the youngest cohort (18-34) who indicated substantially higher WTP in the monthly payment condition. This demographic may find leasing with recurring payments particularly attractive, especially when on average this group face the greatest liquidity constraints for investment in green technologies. Future research could explore whether subsidising a leasing scheme would be a more effective way to increase switching to electric vehicles than the current scheme, where a lump-sum subsidy is given towards a lump-sum investment. A leasing scheme would also spread the risk of technological obsolescence, which might currently be an impediment to purchase. The key point is that when future benefits are undervalued due to some cognitive constraint (limited attention, which causes concentration bias, or underestimation bias, or a combination of both), pricing schemes that reformulate the offer to *balance* the spread of costs and benefits may enhance take-up in a way that improves welfare.

Status Utility and Salience

Chapter 4 illustrated a pattern of car sales in Ireland and Great Britain that is highly unusual by wider international standards, where demand tends to be uniform across the year. The cause of the strong seasonality in car purchase is the age identifier on the licence plate. The age identifier is more salient on the Irish plate. Attributing the greater inequality in sales to the difference in salience was plausible from observation, but made more credible by the causal inference analysis permitted by the switch to bi-annual plates in 2013. The Irish switch to bi-annual plates *increased* rather than decreased the inequality of sales across the year (which was its stated purpose), which is evidence against the maximise-resale-value motive for buying early in a licence plate period. The results seem hard to rationalise without including some status component in the utility function. However, more research is needed to understand what form this status component takes and the contexts in which it receives greater weight in purchase decisions. As mentioned in 6.1, the strength of the inference about the role of status motives - for which the age-identifier is the vessel - would be more credible if supply-side factors could be ruled out. The self-interest of individual sellers to attract consumers may have amplified seasonality in purchases.

The results in Chapter 4 also suggest a promising avenue of theoretical inquiry in strategic games. To date, status signalling has not been modelled using the correlated equilibrium solution concept (Aumann, 1987), in which a choreographer suggests a point of coordination (Gintis, 2014). The point of coordination is essentially a social norm which acts as a focal point (Schelling, 1980). Gintis (2014) has underlined the failure of traditional game theory to explain strategic market phenomena and the need to incorporate the notion of social norms. The salient age-identifier seems an ideal example of a point of coordination. There is potential here to improve on the leading theory of fashion cycles (Pesendorfer, 1995) which contains several unrealistic assumptions according to Coelho et al. (2004). For instance, in this model, the purpose of signalling is to match with potential partners. This means there is no utility derived from simply being seen to possess the status good. Also, when

the fashion seller wants to update their design, they sell off the previous one at cost price. In reality this rarely occurs because brands are wary to maintain their exclusive image. If a status good is to proliferate, it is unlikely to be due to a change in price.

The stylised facts on status and salience in Chapter 4 might be helpful in constructing a more empirically-grounded theory of fashion cycles. One simple change would be to model the variability in the status connotation of a conspicuous consumption good. It seems likely that the salient age identifier is valued highly because its meaning is common knowledge. But the signal of many status goods will be noisier. In theory, those with higher wealth should be more likely to tolerate variance in the status signal to receive a higher expected payoff, because the purchase represents a smaller proportion of their wealth. As early-adopters purchase the noisy status good, it becomes more common, and consequently the variance of its signal falls. This would make it a more attractive purchase to those with lower wealth, despite no change in the price of the status good. This suggests a different mechanism through which fashions propagate. Experimental tests are necessary to test the validity of the assumptions just described.

Future research may investigate the role of simple salient signals of status in other markets. For instance, at the intersection of urban economics and consumer theory, there is scope to investigate whether house prices are disproportionately influenced by postcode. Across cities, postcodes have different levels of salience. For instance, in Dublin there is little disagreement in the ranking of postcodes (generally, lower numbers connote higher status, and even numbers do too). Postcode changes over time, often driven by capacity issues in local sorting offices, provide an opportunity to estimate a postcode premium using a spatial regression discontinuity design which controls for neighbour effects and amenities. [1]

[1] For example, in the 1970s part of D5 was changed to D17. In the 1980s part of D6 became D6W, after residents declined a D12 or D26 postcode. Part of D8 became D10 in the 1980s too.

Group Panels and Decision Quality

A traditional definition of economics is that it is the study of how to allocate scarce resources. Chapter 5 analysed how expert judges weight book attributes when allocating the scarce resource of literary prizes. This may seem like an esoteric topic, but other resources are allocated in a similar way; for instance university places and jobs are often allocated by a qualitative group judgment. These decisions require the integration of multiple incommensurate attributes to arrive at a final ranking of suitability or merit. It is important to understand how decision making processes interact with the environment, and similarly how the proximate goals of decision makers are influenced by the setting. The observational nature of the data in Chapter 5 disallows any claims of causality. Research on accountability and consensus-seeking provided suggestive evidence that part of the bias against shorter novels might stem from a desire to be well-regarded by one's peers in a setting where individual effort is not observable (each takes it on trust that the others have read all the shortlisted novels carefully). Quantitative attributes receive greater weight under these conditions, even if it is not warranted from a normative point of view. Does a similar bias play out in decisions of greater economic consequence? For instance, in hiring decisions, years of experience is a quantitative metric that might be over-weighted. In a group setting, this could generate a bias against otherwise equally qualified candidates who have fewer years experience but are superior on other metrics. These mechanisms warrant further research. Research into decision quality by groups is particularly timely when algorithms are figuratively knocking on the door. Evidence of bias by groups could inform the cost-benefit analysis of devolving some decision-making capacity to algorithms (or alternatively, to appropriately incentivised anonymous crowds).

Concluding Thoughts

My hope is that the research contained in this thesis will make a modest contribution to the ongoing reconvergence of psychology with economics. A core insight from psychology, summarily defenestrated in the Walrasian turn (Bowles and Gintis, 2000), is that behaviour and the capability to optimise is context-specific. People tend to be naturally good at tasks that were important for survival. For instance, the mental number line is tuned to be most precise over the numerosity range relevant to survival decisions: Which branch has more fruit? Can I fight off this many lions, or should I run? These pivotal numbers tended to be closer to zero than one hundred, and so natural selection ensured processing power was deployed accordingly (Barlow et al., 1961). The predictive power of economic models - and their validity as guides to policy - may be enhanced by incorporating the 'number sense' to understand consumption choices that may be susceptible to underestimation bias. A secondary theme that emerged in the latter half of the thesis is that social concerns may *moderate* the decision weight given to numeric or quantitative attributes. However, the observational nature of the data in these chapters means this relationship is speculative. Scientific examination of decision-making processes with a social dimension may produce insight into how people allocate their attention - and in consumption decisions specifically, why they consequently expend their budgets as they do. Insights of this nature may be particularly informative for regulators who have the power to alter the social context of decision-making through the design of policies and institutions.

Bibliography

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abeler, J. and Marklein, F. (2017). Fungibility, labels, and consumption. *Journal of the European Economic Association*, 15(1):99–127.

Adams, G. S., Converse, B. A., Hales, A. H., and Klotz, L. E. (2021). People systematically overlook subtractive changes. *Nature*, 592(7853):258–261.

Agarwal, S., Qian, W., and Zou, X. (2020). Thy neighbor's misfortune: Peer effect on consumption. *American Economic Journal: Economic Policy, forthcoming*.

Alekseev, A., Charness, G., and Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, 134:48–59.

Allcott, H. (2016). Paternalism and Energy Efficiency: An overview. *Annual Review of Economics*, 8:145–176.

Allcott, H. and Knittel, C. (2019). Are consumers poorly informed about fuel economy? evidence from two experiments. *American Economic Journal: Economic Policy*, 11(1):1–37.

Allcott, H. and Wozny, N. (2014). Gasoline prices, fuel economy, and the energy paradox. *Review of Economics and Statistics*, 96(5):779–795.

Andor, M. A., Gerster, A., and Sommer, S. (2020). Consumer inattention, heuristic thinking and the role of energy labels. *The Energy Journal*, 41(1).

Anobile, G., Cicchini, G. M., and Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. *Cognition*, 122(3):454–459.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52(1):388–407.

Ashworth, J., Heyndels, B., and Werck, K. (2010). Expert judgements and the demand for novels in Flanders. *Journal of Cultural Economics*, 34(3):197–218.

Attari, S. Z., DeKay, M. L., Davidson, C. I., and De Bruin, W. B. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of sciences*, 107(37):16054–16059.

Aumann, R. J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18.

Bagwell, L. S. and Bernheim, B. D. (1996). Veblen effects in a theory of conspicuous consumption. *The American Economic Review*, pages 349–373.

Barber, B. M., Heath, C., and Odean, T. (2003). Good reasons sell: Reason-based choice among group and individual investors in the stock market. *Management Science*, 49(12):1636–1652.

Barlow, H. B. et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, 1(01).

Benartzi, S. and Thaler, R. (2007). Heuristics and biases in retirement savings behavior. *Journal of Economic Perspectives*, 21(3):81–104.

Berkouwer, S. B. and Dean, J. T. (2019). Credit and attention in the adoption of profitable energy efficient technologies in Kenya.

Bernheim, B. D. (2009). Behavioral welfare economics. *Journal of the European Economic Association*, 7(2-3):267–319.

Bertrand, M. and Morse, A. (2016). Trickle-down consumption. *Review of Economics and Statistics*, 98(5):863–879.

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, 92(8-9):1787–1794.

Blasch, J., Filippini, M., and Kumar, N. (2019). Boundedly rational consumers, energy and investment literacy, and the display of information on household appliances. *Resource and Energy Economics*, 56:39–58.

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

Bordalo, P., Gennaioli, N., and Shleifer, A. (2016). Competition for attention. *The Review of Economic Studies*, 83(2):481–513.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2021). Salience. Technical report, National Bureau of Economic Research.

Bordalo, P., Gennaioli, N., Shleifer, A., et al. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.

Bowles, S. and Gintis, H. (2000). Walrasian economics in retrospect. *The quarterly journal of economics*, 115(4):1411–1439.

Bradford, D., Courtemanche, C., Heutel, G., McAlvanah, P., and Ruhm, C. (2017). Time preferences and consumer behavior. *Journal of Risk and Uncertainty*, 55(2):119–145.

Brezis, N., Bronfman, Z. Z., and Usher, M. (2015). Adaptive spontaneous transitions between two mechanisms of numerical averaging. *Scientific Reports*, 5:10415.

Bursztyn, L., Ferman, B., Fiorin, S., Kanz, M., and Rao, G. (2018). Status goods: Experimental evidence from platinum credit cards. *The Quarterly Journal of Economics*, 133(3):1561–1595.

Bushong, B., Rabin, M., and Schwartzstein, J. (2021). A model of relative thinking. *The Review of Economic Studies*, 88(1):162–191.

Busse, M. R., Lacetera, N., Pope, D. G., Silva-Risso, J., and Sydnor, J. R. (2013). Estimating the effect of salience in wholesale and retail car markets. *American Economic Review*, 103(3):575–79.

Camerer, C. F. and Johnson, E. J. (1991). The process-performance paradox in expert judgment-how can experts know so much and predict so badly?

Carroll, J., Denny, E., and Lyons, S. (2016). The effects of energy cost labelling on appliance purchasing decisions: Trial results from ireland. *Journal of Consumer Policy*, 39(1):23–40.

Cason, T. N., Masters, W. A., and Sheremeta, R. M. (2020). Winner-take-all and proportional-prize contests: theory and experimental results. *Journal of Economic Behavior & Organization*, 175:314–327.

Charles, K. K., Hurst, E., and Roussanov, N. (2009). Conspicuous consumption and race. *The Quarterly Journal of Economics*, 124(2):425–467.

Coelho, P. R., Klein, D. B., and McClure, J. E. (2004). Fashion cycles in economics. *Econ Journal Watch*, 1(3):437.

Criscuolo, P., Dahlander, L., Grohsjean, T., and Salter, A. (2021). The sequence effect in panel decisions: Evidence from the evaluation of research and development projects. *Organization Science*, 32(4):987–1008.

Cubitt, R. P., Munro, A., and Starmer, C. (2004). Testing explanations of preference reversal. *The Economic Journal*, 114(497):709–726.

Dehaene, S. et al. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. *Sensorimotor foundations of higher cognition*, 22:527–574.

Dehaene, S., Izard, V., Spelke, E., and Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in western and amazonian indigene cultures. *Science*, 320(5880):1217–1220.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic literature*, 47(2):315–72.

DellaVigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review*, 96(3):694–719.

Dertwinkel-Kalt, M., Gerhardt, H., Riener, G., Schwerter, F., and Strang, L. (2022). Concentration bias in intertemporal choice. *The Review of Economic Studies*, 89(3):1314–1334.

English, J. F. (2014). The economics of cultural awards. In *Handbook of the Economics of Art and Culture*, volume 2, pages 119–143. Elsevier.

Englmaier, F., Schmöller, A., and Stowasser, T. (2018). Price discontinuities in an online market for used cars. *Management Science*, 64(6):2754–2766.

Enke, B. (2020). What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Epley, N. and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40.

Fechner, G. T. (1948). Elements of psychophysics, 1860.

Frank, R. H. (1985). The demand for unobservable and other nonpositional goods. *The American Economic Review*, 75(1):101–116.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401.

Frey, B. S. and Gallus, J. (2017). Towards an economics of awards. *Journal of Economic Surveys*, 31(1):190–200.

Fuentealba, C. L. D., Mendoza, J. A. M., Yelpo, S. M. S., Ramos, C. L. V., and Fuentes-Solís, R. A. (2021). Household debt, automatic bill payments and inattention: Theory and evidence. *Journal of Economic Psychology*, 85:102385.

Gerarden, T. D., Newell, R. G., and Stavins, R. N. (2017). Assessing the energy-efficiency gap. *Journal of Economic Literature*, 55(4):1486–1525.

Gilbert, B. and Zivin, J. G. (2014). Dynamic salience with intermittent billing: Evidence from smart electricity meters. *Journal of Economic Behavior & Organization*, 107:176–190.

Gillingham, K. and Palmer, K. (2020). Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. *Review of Environmental Economics and Policy*.

Gillingham, K. T., Houde, S., and Van Benthem, A. A. (2021). Consumer myopia in vehicle purchases: evidence from a natural experiment. *American Economic Journal: Economic Policy*, 13(3):207–38.

Ginsburgh, V. (2003). Awards, success and aesthetic quality in the arts. *Journal of Economic Perspectives*, 17(2):99–111.

Ginsburgh, V. and Weyers, S. (2014). Nominees, winners, and losers. *Journal of Cultural Economics*, 38(4):291–313.

Ginsburgh, V. A. and Van Ours, J. C. (2003). Expert opinion and compensation: Evidence from a musical competition. *American Economic Review*, 93(1):289–296.

Gintis, H. (2014). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences-Revised Edition*. Princeton University Press.

Gourville, J. T. (1998). Pennies-a-day: The effect of temporal reframing on transaction evaluation. *Journal of Consumer Research*, 24(4):395–408.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Greene, D. L. (2010). How consumers value fuel economy: A literature review.

Gupta, N., Rigott, L., and Wilson, A. (2021). The experimenters' dilemma: Inferential preferences over populations. Technical report, arXiv. org.

Hausman, J. A. (1979). Individual discount rates and the purchase and utilization of energy-using durables. *The Bell Journal of Economics*, pages 33–54.

Heinzle, S. L. (2012). Disclosure of energy operating cost information: A silver bullet for overcoming the energy-efficiency gap? *Journal of Consumer Policy*, 35(1):43–64.

Hershfield, H. E., Shu, S., and Benartzi, S. (2020). Temporal reframing and participation in a savings program: A field experiment. *Marketing Science*, 39(6):1039–1051.

Hertwig, R. (2015). Decisions from experience. *The Wiley Blackwell Handbook of Judgment and Decision Making*, 2:239–267.

Hertwig, R. (2017). When to consider boosting: some rules for policy-makers. *Behavioural Public Policy*, 1(2):143–161.

Hinterhuber, A. and Liozu, S. M. (2017). Is innovation in pricing your next source of competitive advantage? In *Innovation in Pricing*, pages 11–27. Routledge.

Hsee, C. K., Rottenstreich, Y., and Xiao, Z. (2005). When is more better? On the relationship between magnitude and subjective value. *Current Directions in Psychological Science*, 14(5):234–237.

Imas, A. and Madarász, K. (2020). Mimetic dominance and the economics of exclusion: Private goods in public context.

Ioannidis, J., Stanley, T., and Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, 127(605):F236–F265.

Irwin, J. R. and Davis, J. H. (1995). Choice/matching preference reversals in groups: Consensus processes and justification-based reasoning. *Organizational Behavior and Human Decision Processes*, 64(3):325–339.

Jacobsen, G. D. (2015). Consumers, experts, and online product evaluations: Evidence from the brewing industry. *Journal of Public Economics*, 126:114–123.

Kahneman, D. and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49:81.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454.

Kallbekken, S., Sælen, H., and Hermansen, E. A. (2013). Bridging the energy efficiency gap: A field experiment on lifetime energy costs and household appliances. *Journal of Consumer Policy*, 36(1):1–16.

Klayman, J. and Brown, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition*, 49(1-2):97–122.

Klemperer, P. (1995). Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. *The Review of Economic Studies*, 62(4):515–539.

Kőszegi, B. and Szeidl, A. (2013). A model of focusing in economic choice. *The Quarterly Journal of Economics*, 128(1):53–104.

Kruger, J., Wirtz, D., Van Boven, L., and Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, 40(1):91–98.

Lacetera, N., Pope, D. G., and Sydnor, J. R. (2012). Heuristic thinking and limited attention in the car market. *American Economic Review*, 102(5):2206–36.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478.

Larrick, R. P. and Soll, J. B. (2008). The mpg illusion.

Lee, J. and Hogarth, J. M. (2000). Consumer information search for home mortgages: who, what, how much, and what else? *Financial Services Review*, 9(3):277–293.

Lerner, J. S. and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2):255.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2):153–174.

Levy, M. R. and Tasoff, J. (2017). Exponential-growth bias and overconfidence. *Journal of Economic Psychology*, 58:1–14.

Li, X. and Camerer, C. (2020). Predictable effects of bottom-up visual salience in experimental decisions and games. *Available at SSRN 3308886*.

Lichtenstein, S. and Slovic, P. (2006). *The construction of preference*. Cambridge University Press.

Linzmajer, M., Hubert, M., and Hubert, M. (2021). It's about the process, not the result: An fMRI approach to explore the encoding of explicit and implicit price information. *Journal of Economic Psychology*, page 102403.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109(453):25–34.

Mazumdar, T., Raj, S. P., and Sinha, I. (2005). Reference price research: Review and propositions. *Journal of Marketing*, 69(4):84–102.

McGowan, F. P., Lunn, P., and Robertson, D. A. (2019). Underestimation of money growth and pensions: Experimental investigations. Technical report, ESRI Working Paper no. 611.

Mehta, J., Starmer, C., and Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, 84(3):658–673.

Milkman, K. L., Carmona, R., and Gleason, W. (2007). A statistical analysis of editorial influence and author–character similarities in 1990s New Yorker Fiction. *Literary and Linguistic Computing*, 22(3):305–328.

Milkman, K. L., Payne, J. W., and Soll, J. B. (2015). A user's guide to debiasing. *Wiley Blackwell Handbook of Judgment and Decision Making*.

Min, J., Azevedo, I. L., Michalek, J., and de Bruin, W. B. (2014). Labeling energy cost on light bulbs lowers implicit discount rates. *Ecological Economics*, 97:42–50.

Minson, J. A., Mueller, J. S., and Larrick, R. P. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science*, 64(9):4177–4192.

Mollick, E. and Nanda, R. (2016). Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science*, 62(6):1533–1553.

Morewedge, C. K., Kassam, K. S., Hsee, C. K., and Caruso, E. M. (2009). Duration sensitivity depends on stimulus familiarity. *Journal of Experimental Psychology: General*, 138(2):177.

Moseley, M. et al. (2019). How the Booker Prize Won the Prize. *American, British and Canadian Studies*, (33):206–221.

Murphy, J. J., Allen, P. G., Stevens, T. H., and Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, 30(3):313–325.

National-Book-Awards (2021). National Book Awards website. `https://www.nationalbook.org/national-book-awards/how-works/`. [Online; accessed 20-September-2021].

Newell, R. G. and Siikamäki, J. (2014). Nudging energy efficiency behavior: The role of information labels. *Journal of the Association of Environmental and Resource Economists*, 1(4):555–598.

O'Connell, H. (2012). Healy-Rae vindicated as 2013 licence plates to avoid "Unlucky 13". *thejournal.ie*.

Olschewski, S., Newell, B. R., Oberholzer, Y., and Scheibehenne, B. (2021). Valuation and estimation from experience. *Journal of Behavioral Decision Making*, 34(5):729–741.

Pesendorfer, W. (1995). Design innovation and fashion cycles. *The american economic review*, pages 771–792.

Ponzo, M. and Scoppa, V. (2015). Experts' awards and economic success: Evidence from an Italian literary prize. *Journal of Cultural Economics*, 39(4):341–367.

Prati, A. (2017). Hedonic recall bias: Why you should not ask people how much they earn. *Journal of Economic Behavior & Organization*, 143:78–97.

Quispe-Torreblanca, E. G., Stewart, N., Gathergood, J., and Loewenstein, G. (2019). The red, the black, and the plastic: paying down credit card debt for hotels, not sofas. *Management Science*, 65(11):5392–5410.

Read, D. (2001). Is time-discounting hyperbolic or subadditive? *Journal of Risk and Uncertainty*, 23(1):5–32.

Scheibehenne, B. (2019). The psychophysics of number integration: Evidence from the lab and from the field. *Decision*, 6(1):61.

Schelling, T. C. (1980). *The Strategy of Conflict: With a New Preface by The Author*. Harvard university press.

Schley, D. R. and Peters, E. (2014). Assessing "economic value" symbolic-number mappings predict risky and riskless valuations. *Psychological Science*, 25(3):753–761.

Scholten, M. and Read, D. (2006). Discounting by intervals: A generalized model of intertemporal choice. *Management Science*, 52(9):1424–1436.

Schwarz, N. and Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513.

Sexton, S. (2015). Automatic bill payment and salience effects: Evidence from electricity consumption. *Review of Economics and Statistics*, 97(2):229–241.

Siegler, R. S. and Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, 14(3):237–250.

Silvi, M. and Rosa, E. P. (2021). Reversing impatience: Framing mechanisms to increase the purchase of energy-saving appliances. *Energy Economics*, 103:105563.

Simon, H. A. (1986). Rationality in psychology and economics. *Journal of Business*, pages 209–224.

Slovic, P. (1975). Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3):280.

Söderberg, M. and Barton, D. N. (2014). Marginal WTP and distance decay: The role of 'protest'and 'true zero' responses in the economic valuation of recreational water quality. *Environmental and Resource Economics*, 59(3):389–405.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, 38(2):317–346.

Somerville, J. (2022). Range-dependent attribute weighting in consumer choice: An experimental test. *Econometrica*, 90(2):799–830.

Stango, V. and Zinman, J. (2009). Exponential growth bias and household finance. *The Journal of Finance*, 64(6):2807–2849.

Stango, V. and Zinman, J. (2020). We are all behavioral, more or less: A taxonomy of consumer decision making. Technical report, National Bureau of Economic Research.

Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, 67(2):76–90.

Takahashi, T. (2006). Time-estimation error following Weber–Fechner law may explain subadditive time-discounting. *Medical hypotheses*, 67(6):1372–1374.

Taylor, S. E. and Thompson, S. C. (1982). Stalking the elusive" vividness" effect. *Psychological Review*, 89(2):155.

Tetlock, P. E. (2009). *Expert political judgment*. Princeton University Press.

Tetlock, P. E. and Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of personality and social psychology*, 57(3):388.

Tindale, R. S. and Kluwe, K. (2015). Decision making in groups and organizations. *The Wiley Blackwell handbook of judgment and decision making*, 2:849–874.

Tindale, R. S. and Winget, J. R. (2019). Group decision-making. In *Oxford Research Encyclopedia of Psychology*.

Troutman, C. M. and Shanteau, J. (1976). Do consumers evaluate products by adding or averaging attribute information? *Journal of Consumer Research*, 3(2):101–106.

Tsetsos, K., Chater, N., and Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, 109(24):9659–9664.

Turrentine, T. S. and Kurani, K. S. (2007). Car buyers and fuel economy? *Energy Policy*, 35(2):1213–1223.

Tversky, A. and Thaler, R. H. (1990). Anomalies: preference reversals. *Journal of Economic Perspectives*, 4(2):201–211.

Wichman, C. J. (2017). Information provision and consumer behavior: A natural experiment in billing frequency. *Journal of Public Economics*, 152:13–33.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12:579–601.