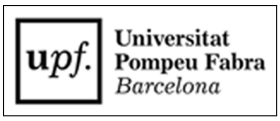
Essays in Behavioural Economics

Katharina A. Janezic

TESI DOCTORAL UPF / year 2021

THESIS SUPERVISORS
Professor Larbi Alaoui, Professor Jose Apesteguia
Department of Economics and Business



To my family.

Acknowledgements

Many people have been instrumental to my journey to this thesis. I would like to take the opportunity to thank them here.

First of all, I would like to extend my heartfelt gratitude to my supervisors Larbi Alaoui and Jose Apesteguia for their guidance, their advice on the chapters contained in this thesis and for their invaluable support, encouragement and mentoring throughout the years.

I also owe special thanks to Antonio Penta for his generous support and advice throughout my studies and for being a wonderful mentor. I would also like to thank my co-author Aina Gallego for our great collaboration. Furthermore, I would like to thank Rosemarie Nagel as well as Gaël Le Mens for their valuable advice, especially on experimental methods.

Many professors have given me their advice and time throughout my Ph.D. whom I would like to thank here: Johannes Abeler, Francesco Cerigioni, George Chondrakis, Libertad González, Pedro Rey Biel. I would also like to thank our wonderful lab manager Pablo López-Aguilar and Marta Araque and Laura Agustí for their outstanding administrative support.

I received very useful comments from seminar participants at the 2020 Econometric Society European Winter Meetings, the 2020 Barcelona GSE PhD Jamboree, the Annual Meeting of the International Society of Political Psychology 2019, at Universitat Pompeu Fabra, London School of Economics, Université Catholique Louvain la Neuve and Paris School of Economics.

In addition, I would like to thank my colleagues for our numerous coffee breaks and many interesting lunch-time discussions throughout our shared years in Barcelona: Adam, Ana, André, Bjarni, Christoph, Daniel, David, Karolis, Madalen, Sergio and Yiru.

I would like to thank my wonderful family, Mum, Dad, Steffi and Ben, for their unwavering love and support throughout my entire life, for encouraging me to achieve my dreams and to believe in myself.

Thank you for always being interested in hearing about my thesis, for providing perspective and for always being there for me.

Finally, I would like to thank my partner Lukas for going on this journey together, for making me smile throughout it and for always being there to discuss any aspect of my research. Thank you for your constant support and love.

Thank you so much to all of you.

Abstract

This thesis contains three chapters studying questions of behavioural and experimental economics.

Chapter 1, titled "Heterogeneity in lies and lying preferences", develops a theoretical framework and an experimental design which I use to identify systematic patterns of lying behaviour in the presence of heterogeneity of lies and decision-makers. I show that accounting for these patterns provides large gains in out-of-sample predictions of lying decisions.

Chapter 2, titled "Eliciting preferences for truth-telling in a sample of politicians", studies the connection between politicians' truth-telling preferences and observable variables such as re-election success. The chapter has been published in the *Proceedings of the National Academy of Sciences*.

Chapter 3, titled "Reasoning about others' reasoning", introduces an experimental design strategy to disentangle cognitive from belief-based levels of play in a model of iterative reasoning based on observed choices in an experiment. The chapter has been published in the *Journal of Economic Theory*.

Resum

Aquesta tesi conté tres capítols que estudien qüestions d'economia del comportament i experimental.

El capítol 1, titulat "Heterogeneïtat en mentides i preferències per mentir", desenvolupa un marc teòric i un disseny experimental que faig servir per identificar patrons sistemàtics de comportament mentider en presència d'heterogeneïtat de mentides i de decisors. Demostro que tenir en compte aquests patrons proporciona grans guanys en prediccions fora de la mostra.

El capítol 2, titulat "Obtenir preferències per dir la veritat en una mostra de polítics", estudia la connexió entre les preferències per dir la veritat dels polítics i variables observables com l'èxit de la reelecció. El capítol s'ha publicat a *Proceedings of the National Academy of Sciences*.

El capítol 3, titulat "Raonar sobre el raonament dels altres", introdueix una estratègia de disseny experimental per diferenciar en un joc els nivells d'actuació basats en creences dels basats en límits cognitius en un model de raonament iteratiu basat en les decisions observades en un experiment. El capítol s'ha publicat al *Journal of Economic Theory*.

Preface

This thesis consists of three chapters on topics in behavioural and experimental economics.

In Chapter 1, titled "Heterogeneity in lies and lying preferences", I examine individual-level lying preferences in the presence of multiple types of lies. I provide a unifying framework to analyse and predict lying behaviour in the context of heterogeneous decision-makers and lies. Lie types are defined by the consequences that they have for a decision-maker and another passive individual who is affected by the decision-maker's choices. The paper has a two-pronged approach in that it provides both a theoretical framework with testable predictions and a novel experimental design. My design allows the observation of individual choices, enables to assess variation of behaviour across lie types as well as to disentangle the effects of lie types from standard social preferences. I find systematic patterns of behaviour in the presence of the different lie types. I analyse and predict lying behaviour non-parametrically using an unsupervised machine learning algorithm uniquely suited to classification. I also employ a parametric approach in which I introduce a parametric version of the theoretical framework and estimate it using maximum likelihood. In a further analysis, I directly contrast social preferences with lying preferences, which were both elicited in the experiment. I find that social preferences have only limited explanatory power for lying preferences, as measured by their out-of-sample predictability. Finally, I show that accounting for the identified systematic patterns of lying behaviour leads to large out-of-sample forecasting gains.

In Chapter 2, titled "Eliciting preferences for truth-telling in a survey of politicians", which is joint work with Aina Gallego, we examine honesty and its possible correlates in a sample of politicians. In this paper, we elicit politicians' willingness to lie using an experiment with a novel incentivisation mechanism. We then examine whether sub-groups are heterogeneous with regard to their propensity to lie and whether these differences have an impact on political outcomes.

In order to measure these preferences, we collected a large sample of 816 survey responses from elected politicians from Spain, a country that is representative of advanced industrialised democracies. We first show that a significant percentage of politicians lied in our survey. Contrary to popular opinion, we find no difference in the lying propensity of male and female mayors but show that members of the two largest parties lie significantly more than those of smaller parties. We find that dishonesty is significantly and positively correlated with re-election success, even when accounting for possible confounds such as the likelihood of re-running for office. The paper has been published in the *Proceedings of the National Academy of Sciences* (Janezic and Gallego (2020)). The order of the paper's sections has been retained from the published version.

Chapter 3, titled "Reasoning about others' reasoning", which is joint work with Larbi Alaoui and Antonio Penta, introduces a novel experimental design strategy aimed at disentangling whether an observed level of play in a level-k setting of iterative reasoning was determined by cognitive ability or by beliefs. In the paper, we formalise predictions and the corresponding identification assumptions based on the endogenous depth of reasoning model (EDR, Alaoui and Penta (2016a)) and test them in a series of experiments. We present two paradigms, the replacement and the tutorial methods, to allow the disentanglement of cognitive from behavioural levels. The replacement method serves to remove beliefs of second or higher order by ensuring that a subject's opponent is not playing against the subject but against a third person. This ensures that the subject does not have to anticipate what the opponent thinks about the subject's potential level of play. The tutorial method removes cognitive constraints by explaining the chain of reasoning to the subjects. This allows to assess whether a choice in the experiment was driven by beliefs or by cognitive ability. For example, if there are no observed changes in behaviour after removing the cost of reasoning (via the tutorial method), while keeping beliefs about the opponent fixed (via the replacement method), this implies that the original level of play was

determined by beliefs about the opponent's ability. We employ the paradigms in three different experiments to show that thinking about others' reasoning is more nuanced than previously assumed in the literature. In particular, we show that levels of play can indeed be determined either by a subject's own cognitive ability or by beliefs about others' cognitive ability. The paper has been published in the *Journal of Economic Theory* (Alaoui et al. (2020)).

Contents

List of figures **xx**

List of tables **xxii**

1 HETEROGENEITY IN LIES AND LYING PREFERENCES **1**

1.1 Introduction 1
1.1.1 Literature Review 6
1.2 Theoretical Framework 9
1.3 Experimental Design and Logistics 16
1.3.1 Logistics 16
1.3.2 Design 19
1.3.2.1 Lying game 19
1.3.2.2 Social preference game 25
1.4 Results of the Lying Game 26
1.4.1 Aggregate results 26
1.4.2 Do lie types matter empirically? 28
1.4.3 Heterogeneity of preferences across individuals 31
1.5 Lies and social preferences 41
1.5.1 Are behavioural groups in the lying game comparable to those in the social preference game? 42
1.5.2 Individual specific analysis 48
1.6 Parametrised and Calibrated Utility Function 50
1.6.1 Utility Function 50

1.6.2 Calibration 52
1.7 Conclusion 57
1.8 Appendix 59
1.8.1 Instructions of the Experiment 59
1.8.1.1 Introduction 59
1.8.1.2 Instructions - Lying game 60
1.8.1.3 Instructions - Social preference game 63
1.8.2 Specification of rounds of the experiment . . . 65
1.8.3 Details on comparison of aggregate results to the literature 67
1.8.3.1 Lying behaviour across charities . . . 67
1.8.3.2 Covariates of lying 69
1.8.4 Supporting results for Section 1.4 72
1.8.4.1 Robustness to misspecification of number of clusters k 72
1.8.4.2 Never-liars and potential errors of behaviour 72
1.8.4.3 Behaviour of clusters 73
1.8.5 Supporting results for Section 1.5 74
1.8.6 Supporting results for Section 1.6 75
1.8.7 Glossary for machine learning methodology . . 77

2 ELICITING PREFERENCES FOR TRUTH-TELLING IN A SURVEY OF POLITICIANS 79
2.1 Introduction 80
2.2 Results 84
2.2.1 Gender 85
2.2.2 Party membership 87
2.2.3 Reelection 88
2.3 Discussion 92
2.4 Materials and Methods 94
2.4.1 Data Availability 94
2.4.2 Setting, Participants, and Fieldwork 94
2.4.3 New measure of lying 96

2.4.4 Statistical Analyses and Regression Model . . 98
2.4.5 Robustness Analyses 100

3 REASONING ABOUT OTHERS' REASONING 105
3.1 Introduction 106
3.1.1 Related Literature 110
3.2 Baseline Game and General Logistics 113
3.2.1 The acyclical 11-20 game 113
3.2.2 General Logistics 115
3.2.3 Subjects' Classifications 116
3.3 Experimental Design 117
3.3.1 Experiment 1: Treatments 117
3.3.1.1 Baseline Treatments 118
3.3.1.2 Relaxing Cognitive Bounds: the post-Tutorial Treatments 119
3.3.1.3 Reasoning about Others' Incentives: Asymmetric Payoffs Treatments . . . 121
3.3.2 Experiment 2: Unlabeled Variations 123
3.3.3 Experiment 3: Unlabeled Variations with preliminary semi-Tutorial 125
3.4 The EDR model 126
3.4.1 Baseline Model 128
3.4.2 Identification Assumptions and Predictions for the Experiments 132
3.5 Results 139
3.5.1 Experiment 1 140
3.5.1.1 Summary of AP's main results . . . 140
3.5.1.2 Relaxing cognitive bounds – Experimental Results 141
3.5.1.3 Reasoning about opponents' incentives – Experimental Results 146
3.5.2 Experiment 2: Results 150
3.5.3 Experiment 3: Results 152
3.5.4 Individual Effects 156

3.6 Concluding Remarks 158
3.7 Appendix 162
3.7.1 Proofs 162
3.7.2 Logistics of the Experiments 167
3.7.2.1 Instructions of the Experiment 168
3.7.2.2 Sequences 172
3.7.2.3 Details of the Cognitive Test 172
3.7.3 Regressions 176
3.7.4 Additional Figures 183
3.7.4.1 TOM scores 183
3.7.4.2 Individual behavior: violations of theory 184
3.7.4.3 Individual behavior: shifts in behavior 186

List of Figures

1.1 Visualisation of Property 3 14
1.2 Choice of Charities 18
1.3 Example of the 1st screen of one of the rounds of the lying game 20
1.4 Example of the 2nd screen of one of the rounds of the lying game 21
1.5 Payoff space for binary questions 24
1.6 Example screen of one of the rounds of the social preference game 26
1.7 Percentage of liars by lie type 27
1.8 Scree plot of the 10 largest components in lying game PCA analysis 34
1.9 Performance indicators by cluster number in lying game k-means analysis 35
1.10 Representative DMs by behavioural lying cluster . . . 36
1.11 Out-of-sample performance without and with heterogeneity in preferences taken into account 40
1.12 Scree plot of 20 largest components in social preference game PCA analysis 42
1.13 Representative DMs by behavioural social preference cluster 44
1.14 Performance of predicting lying behaviour based on social preference behaviour 47
1.15 Estimated lie regions for the representative agent of the lying game 52

1.16 Estimated lie regions for representative agents of groups identified in Section 1.4 56
1.17 Example screen of the 1st screen of one of the rounds of the lying game. 60
1.18 Example screen of the 2nd screen of one of the rounds of the lying game. 61
1.19 Example screen of one of the rounds of the social preference game. 63
1.20 Average percentage of lies by lie type for each charity 68
1.21 Histogram of each subject's MSE in out-of-sample prediction exercise with heterogeneity in preferences taken into account for different k. 72
1.22 Percentage of lies told for each question in the lying game by cluster 73
1.23 Performance indicators by cluster number in social preference game k-means analysis 74
1.24 Estimated lie regions for representative agents based on the mean behaviour of each cluster identified in Section 1.4 76

2.1 **Proportion of mayors who report heads and tails and interest in receiving the report**. (A) The percentage of mayors reporting heads and tails is displayed above the bars, showing that they differ significantly from the objective 50% benchmark (two-sided binomial test), indicating a high frequency of lying. (B) The percentage of mayors who reported heads depending on their interest in the report is given by the height of the bars and additionally, at the bottom of the bars, the share of mayors in each category of interest in the report is displayed. Standard errors around the mean are given by the intervals. Stars indicate a significant deviation from the 50% benchmark calculated by a two-sided binomial test, *** p-value<0.01. 83

2.2 **Percentage of mayors who reported heads by their individual characteristics**. (A) The percentage of mayors reporting heads by their gender and (B) membership in a large party. All categories exceed the 50% benchmark (two-sided binomial test). The difference between genders is negligible. However, there is a highly significant difference between members from major versus minor political parties. Standard errors are given as intervals around the mean. Stars indicate significant deviation of reported heads between subgroups (two-sample t test), ** p-value< 0.05 . . . 86

2.3 **Percentage of mayors who reported heads by measures relevant to reelection**. The percentage of mayors reporting heads by (A) their reported desire to rerun for office, (B) actually rerunning for office, and (C) reelection results. All categories exceed the 50% benchmark (two-sided binomial test). The difference between those who want to rerun and those who do not is small and not statistically significant. Similarly, there is no statistical difference between those who reran for office and those who did not. However, there is a highly significant difference between reelected and not reelected mayors. Standard errors are given as intervals around the mean. Stars indicate significant deviation of reported heads between subgroups (two-sample t-test), ** p-value< 0.05 89

3.1 Pre- and post-tutorial comparisons, label I (top) and label II (bottom) 142

3.2 Beliefs comparisons: post tutorial treatments, label I (left) and label II (right) 145

3.3 Asymmetric payoffs treatments, label I (left) and label II (right). 146

3.4 Asymmetric payoffs treatments, with double replacement. Label I (top) and label II (bottom). 148
3.5 Beliefs comparisons: asymmetric payoffs treatments, label I (left) and label II (right). 149
3.6 Pre- and post-tutorial, unlabeled treatments. 150
3.7 Asymmetric payoffs comparisons, unlabeled treatments. 151
3.8 Pre- and post-tutorial, unlabeled treatments, Experiment 3. 153
3.9 Asymmetric payoffs comparisons, unlabeled treatments, Experiment 3. 153
3.10 Distribution of total TOM scores. 155
3.11 Distribution of factual TOM scores. Vertical red line indicates sample average. 183
3.12 Distribution of TOM mind scores. Vertical red line indicates sample average. 183
3.13 Experiment 1 (tutorial sessions): Number of violations of theory for averaged treatments. 184
3.14 Experiment 1 (payoff sessions): Number of violations of theory for averaged treatments. 184
3.15 Experiment 2: Number of violations of theory for averaged treatments. 185
3.16 Experiment 3: Number of violations of theory for averaged treatments. 185
3.17 Experiment 1: Frequency of shifts in level played from [HOB] to [AP-Het]. 186
3.18 Experiment 3: Frequency of shifts in level played from [Un] to [AT-Un] (left) and from [Un] to [Un+] (right). 187
3.19 Experiment 3: Frequency of shifts in level played from [Un] to [AP-Un] (left) and from [AP-Un] to [Un+] (right). 187

List of Tables

1.1 Lie Types 10
1.2 Percentage of liars by lie type: Comparison with literature 29
1.3 In-sample prediction results 32
1.4 Parameter values by representative agent for the full sample and by group 54
1.5 Rounds with non-binary choice 65
1.6 Significant differences in group behaviour by covariates of interest 71

2.1 Linear probability regressions with gender and membership in a major party as independent variables. . . 102
2.2 Linear probability regressions with reelection as dependent variable. 103

3.1 Summary of the baseline treatments 118
3.2 Summary of the post-tutorial treatments. 120
3.3 Summary of the asymmetric payoff treatments. . . . 122
3.4 Summary of the treatments in Experiment 2. 124
3.5 Summary of the treatments in Experiment 3. 126
3.6 Summary of all treatments over all experiments. . . . 127
3.7 For any subject $t_i = (c_i, c_j^i, c_i^{ij})$, the table shows the classes of treatments that generate the same c_i, c_j^i or c_i^{ij} in Experiments 2 and 3 133

3.8 For any subject $t_i = (c_i, c_j^i, c_i^{ij})$, the table shows the classes of treatments that generate the same c_i, c_j^i or c_i^{ij} in Experiment 1 133
3.9 Summary of results over all experiments. 156
3.10 Experiment 1, Regressions on Payoffs Effects (joint for all sequences) 177
3.11 Experiment 1, Regressions from Post-tutorial treatments. 178
3.12 Experiment 1, Regressions for asymmetric payoff treatments. 178
3.13 Experiment 1, regressions to examine effect of labels for going "From treatment x to y" (using a dummy for Label I, treatment dummy, and a label-treatment interaction term). 179
3.14 Experiment 2, regressions for all treatments. 180
3.15 Experiment 3, regressions for all treatments. 181
3.16 Experiments 1-3, Wilcoxon signed rank test results for all treatments. 182

Chapter 1

HETEROGENEITY IN LIES AND LYING PREFERENCES

1.1 Introduction

For centuries, philosophers and lawmakers have concerned themselves with the concept of lying, debating questions such as whether people should lie, which lies ought to be punishable and how they can be prevented. More recently, economists have joined the debate (see for example Gneezy (2005), Abeler et al. (2019)). Despite this continued interest, we still cannot anticipate when someone will lie, often because individuals' preferences, as well as the lies themselves, differ substantially from each other. How can we take heterogeneity of both lies and lying preferences into account when analysing and predicting individual lying decisions?

In this paper, I provide a unifying framework to analyse individual lying preferences. Previous papers have shown that people are highly heterogeneous when it comes to their lying behaviour (see for example Gibson et al. (2013)). At the same time, most existing research has used methods that analyse lying decisions at the aggregate level

or has elicited individual preferences in environments where they could not be separated from belief effects. Due to the existence of heterogeneity, analysing lying decisions at the aggregate level might come at the cost of individual differences averaging each other out and preventing us from understanding lying behaviour. Similarly, if beliefs about a possible response to (lying) decisions and lying preferences are conflated, it is challenging to elicit lying preferences from observed behaviour. For this reason, it is crucial to research lying preferences at the individual level without confounding beliefs. Moreover, not only the decision-makers are heterogeneous, but so are the lies themselves. Lies differ with respect to their consequences and can be classified into types accordingly. My framework provides a toolkit to help us understand how these heterogeneities interact. The theoretical framework provides a classification of lies as well as a simple model with testable predictions for individual behaviour. At the experimental level, I introduce a novel experimental design that makes it possible to observe lies at the individual level while removing confounding factors such as first or higher order beliefs. In addition to controlling for such beliefs, my experimental design allows the separation of lying preferences from social preferences. Finally, to show that the framework can be used for the purpose of model building and prediction, I present a parametric version of the framework and calibrate it at the individual and the group level.

My unifying framework contributes to the literature along four dimensions. First, at the theoretical level, the framework provides a classification of the lie type space and a simple non-parametric model. Formally, lie types in this paper are defined in terms of lies' consequences on the monetary payoff of a decision-maker and another person who is affected by the lie, which follows the approach first introduced by Gneezy (2005). A complete characterisation is a natural prerequisite for modelling heterogeneity of lies in a systematic manner. The model provides guidance on how to examine behaviour in the context of these lies. It is a private information model with the following set-up: A decision-maker privately observes the true

state of the world and is asked to report it. She can report any state from a set of potential states. The decision-maker and another person, the 'partner', are affected by the reported state in that they receive monetary payoffs connected to that state. There are no strategic concerns, as the partner cannot influence the outcome and the report is unverifiable. Preferences are heterogeneous with respect to the importance that the decision-makers attach to their own and the partner's payoffs as well as with respect to the cost of lying. Behaviour is characterised by a single crossing property: if a subject preferred to lie, she should not revert back to telling the truth when either of the payoffs associated with lying increases relative to the observed lie. This set-up allows the clear definition and systematic variation of lie types needed to study and model their impact on lying behaviour and provides testable predictions that help guide the experimental design. Due to the framework's properties, 'regions of inference' can be constructed in which the decision-maker is expected to lie. These regions provide a non-parametric tool to predict behaviour and can be used to falsify the model.

Second, the paper contributes along the experimental dimension by introducing a novel experimental design that allows the researcher to observe individual choices without suffering from the confounding factors of first or higher order beliefs. It thereby combines the advantages from the two major experimental designs used in the lying literature, die rolling and sender-receiver games, while avoiding their drawbacks. Strategic beliefs make it difficult to elicit lying preferences from observed decisions for the following reason: if a decision-maker decides between telling the truth or a lie to another person and this other person does not know whether it was the truth but can react to it, then expectations of the other person's response will feed back into the decision-maker's decision of whether to lie or not. For example, if she thinks that the other person will not trust her even if she tells the truth, she might tell the truth not because of lying preferences but because of strategic considerations. Eliciting lying preferences from observed behaviour is then very challenging. That beliefs are

inconsequential in my setting in conjunction with knowledge of individual choices makes it possible to pinpoint the effect of lie types on behaviour. The behaviour elicited in the experiment is used to examine the descriptive and predictive power of the framework. The experimental results show that varying the type of lie has a large and significant impact on subjects' behaviour. Moreover, this effect goes beyond differences in payoffs which suggests that the lie type plays a psychological role in itself. Importantly, accounting for lie types improves the performance of prediction exercises, as measured by a reduction in the mean squared forecast error (MSE), by more than 21.5 percent. The results demonstrate the importance of accounting for heterogeneity of lie types in models and in experiments. I show that the impact of lie types is highly heterogeneous across subjects and several behavioural types can be identified in the data. Specifically, I employ a machine learning approach which combines a principal component analysis with a k-means analysis in order to group subjects based on their decisions throughout the experiment. The k-means clustering analysis is a statistical pattern recognition tool that, in economics, is commonly used in time-series econometrics (see for example Falat and Pancikova (2015), Bagnall et al. (2003), Focardi and Fabozzi (2001)). The algorithm identifies six separate behavioural groups which differ with respect to the number and type of lies that they tell in the experiment. Not only can the method uncover heterogeneity but also quantify it. I show that the types of behaviour identified by the algorithm are systematic and meaningful in the sense that subjects within groups make similar choices to each other and that a narrative explaining those choices can be found. Exploiting the uncovered heterogeneity improves out-of-sample forecasting accuracy by more than 60 percent.

Third, the experimental design makes it possible to separate lying preferences from the potential confound of standard social preferences, such as altruism or inequality aversion. In principle, differences in behaviour across lie types could be fully explained by social preferences as lie types are categorised based on variations in monetary outcomes.

If that were the case, we should expect that social preferences can predict how subjects behave in the lying game and a novel lying framework would not be necessary as we could simply rely on the social preference literature to explain lying behaviour in the context of lie types. In order to directly contrast lying preferences with social preferences, the experiment contains two separate but nearly identical games. In the first, called the lying game, a large number of lie-truth choices per individual is elicited. In the second game, subjects choose between the same alternatives as in the lying game but alternatives are no longer classified as truths or lies as there is no truth benchmark. My contribution here is the direct comparison between lying preferences and social preferences. This is possible due to choices being observable at the individual level in the lying game and in the corresponding social preference game. I find that knowing how subjects respond in a social preference game does not help to predict behaviour in the lie setting despite both choice settings being identical to each other except for the choice objects' labels (unlabelled versus truth and lie labels). Thus, the finding confirms that a lying framework is indeed necessary to describe lying preferences even in situations where lie types are fully captured by payoff differences affecting the decision-maker and another person. This further confirms that the effect of lie types goes beyond that of payoff consequences and that they have an additional psychological impact on the decision-maker.

Fourth, having shown that, first, lie types matter and second, that people respond systematically to these lie types, I provide a parametric model that can capture these lying preferences. Specifically, I propose a simple parametric version of the general model that performs well for the data generated by the experiment. Based on the behaviour elicited in the experiment, I calibrate the model's parameters using a maximum likelihood estimation (MLE) approach. I find substantial variation in the estimated lie regions across subjects, in line with the hypothesised decision-maker types.

The theoretical and experimental results highlight that a unifying framework which can account for heterogeneity of both lies and

decision-maker's preferences vastly improves our ability to capture, understand and, most importantly, predict lying behaviour.

The rest of the paper is structured as follows: The next section discusses the related literature. Section 1.2 presents a model and several properties for lying behaviour under lie types. Section 1.3 introduces the design and the logistics of the experiment. The experimental results are presented in Sections 1.4 and 1.5. Section 1.6 introduces a parametrised utility function in line with the framework and discusses the results of a calibration exercise. Section 1.7 concludes the paper.

1.1.1 Literature Review

In his seminal paper, Gneezy (2005) defines lies based on their monetary consequences and Erat and Gneezy (2012) extend this to a more formal definition of lie types. Specifically, each lie type is defined by two outcomes, the monetary payoff of the decision-maker and that of the partner, relative to the payoffs associated with telling the truth. Since then, the literature has examined many different aspects of lying such as the existence of moral costs of lying (Kartik (2009), Gibson et al. (2013)), measures of lying (Fischbacher and Föllmi-Heusi (2013), Gneezy et al. (2013)), or the role of deliberation time (Capraro (2017), Lohse et al. (2018)). While these studies have focused on different explanations of lying preferences, they have in common that they focus their attention on egoistic lies i.e. those lies that benefit the liar at the expense of someone else.

A smaller literature also examines other types of lies such as self-serving (only affect the liar) or mutually beneficial lies (beneficial to all people affected by the lie), and a handful have considered multiple types of lies simultaneously (Erat and Gneezy (2012), Levine and Schweitzer (2014), Biziou-van Pol et al. (2015)). Erat and Gneezy (2012) establish that behaviour varies across the lie types studied. Biziou-van Pol et al. (2015) study both altruistic lies (lies where the

decision-maker incurs a loss and the partner a gain relative to the truth) and mutually beneficial lies. A puzzle in this literature has been the large differences in the percentages of liars across papers. Even in papers that compare lies of the same lie types, this puzzle persists (compare for example Biziou-van Pol et al. (2015), Erat and Gneezy (2012) and Hurkens and Kartik (2009)). I contribute to this literature by utilising a large set of lies in the experimental set-up. While most papers use one to five lie specifications when examining lying behaviour, this paper uses 60. I can thereby systematically assess how behaviour varies across the complete space of lies and show that the type but also the specifics of each lie matter in determining how many people will tell that lie. This underlines the need for systematic analyses as compared to standard practices where one lie is taken as representative of the whole space of lies.

Two experimental paradigms are particularly prominent in the analysis of lying preferences: the die rolling paradigm (Fischbacher and Föllmi-Heusi (2013)) and sender-receiver games (Gneezy (2005), Hurkens and Kartik (2009)). In the die rolling experiments, subjects are asked to privately roll a die and to report the outcome. Monetary payoffs are increasing in the outcome of the die so that subjects have an incentive to report higher numbers irrespective of the true outcome. They thus have an incentive to lie. The empirical distribution of reported outcomes can then be statistically compared to the theoretical distribution which, under the assumption that no one lies, predicts that in sufficiently large samples, all six numbers come up with equal probability. If there exists a statistically significant difference between the theoretical and the empirical distribution, this can be attributed to lying. A popular variation of this methodology is the coin flipping paradigm (see for example Abeler et al. (2014)), where a coin is flipped instead. The die rolling paradigm captures lying preferences at the group level and is belief free in the sense that there are no reputation concerns and no one other than the decision-maker is directly affected by the lie. On the other hand, sender-receiver games can measure individual preferences but introduce confounds via first or higher

order beliefs. Here, a sender decides whether to send a truthful or a dishonest message to the receiver. The receiver then responds with an action, based on beliefs of whether the message was truthful or not. Senders anticipate the response and thus include beliefs about the possible responses in their initial decision. My experiment contributes to the literature by being able to combine the advantages of both of these experimental designs. It captures individual preferences without suffering from the confound of belief effects.

Kerschbamer et al. (2019) as well as Hurkens and Kartik (2009) and Biziou-van Pol et al. (2015) study the connection between social and lying preferences. Most papers that examine this potential connection elicit social preferences in a setting different from the lying experiment that they employ. For example, Kerschbamer et al. (2019) use the *Equality Equivalence Test* and Biziou-van Pol et al. (2015) use both a dictator and a prisoner's dilemma game in order to elicit social preferences. Their analysis is at the group level. Hurkens and Kartik (2009) employ a sender-receiver game and ask all senders to participate in a dictator game. While they find no statistically significant difference, the authors acknowledge that only 55% of senders believed that the receivers would believe their message so that it is difficult to disentangle trust, strategic and distributional preferences from lying preferences. I contribute here by being able to abstract from such confounding factors and by directly contrasting each individual's lying choices with distributional choices rather than relying on group averages.

Abeler et al. (2019) present an extensive meta study on lying. They provide insights into which concerns enter the decision to lie. As a large number of past papers has employed the die rolling and coin flipping paradigms, the authors measure lying preferences at the group level. I complement their analysis by focusing on the individual level, instead. In addition, I contribute by examining lies other than self-serving lies (those lies that benefit the decision-maker and have no direct effect on someone else) and provide a machine learning approach to assess the heterogeneity of lying preferences.

1.2 Theoretical Framework

I consider a setting with two individuals, where one individual has to choose between telling the truth or lying. The first individual is called the *decision-maker* and the second is called the *partner*. The decision-maker reports a privately observed state; the report can either be truthful or a lie and is non-verifiable. The report has payoff consequences for both individuals. The partner is passive in the sense that he cannot influence the outcome but is affected by the decision that the first individual makes via the payoffs. As this paper focuses on intrinsic motives to lie, there is no strategic interaction between decision-makers and their partners. Therefore, beliefs over the partner's actions are inconsequential for the decision-making process. Furthermore, payoffs will be paid out with certainty and can therefore enter the decision-making process directly rather than in expectation.

I now present a framework and some basic properties that formalise behaviour under such a setting. These properties help explain lying behaviour and inform both the design of the experiment and the analysis of the experiment's results.

A decision-maker (DM) faces a decision problem $d \in \{1, \ldots, D\}$. For each decision problem d, there exists a finite set of potential states of the world $S_d = \{s_d^0, s_d^1, \ldots, s_d^M\}, M \geq 1$, with typical element s_d^m, $m = 0, 1, \ldots, M$. Each decision problem d contains one *true state*, which is denoted by s_d^0. All other states in S_d are *untrue states*. The DM privately observes S_d with true state s_d^0. Any state in S_d can be publicly reported as the true state by the DM, irrespective of whether it is the true state s_d^0 or one of the untrue states. The states are payoff relevant in that any reported state s_d^m is tied to monetary payoffs $(x_d^m, y_d^m) \in \mathbb{R}_+^2$. The DM's payoff is given by x_d^m and the partner's payoff is y_d^m. The DM knows the mapping between states and payoffs. The DM chooses to report the state that maximises his utility, which depends on both the true and the reported state. The state that is reported is denoted by r_d. We say that a DM decides to lie whenever $r_d \neq s_d^0$.

Table 1.1 defines the possible lie types when s_d^0 is the true state and s_d^m is any other potential state. Lie types are defined by comparing the decision-maker's and the partner's payoffs from lying relative to telling the truth. To illustrate, imagine that a DM faces a decision problem d where the true state s_d^0 is associated with the payoff bundle $(x_d^0, y_d^0) = (5, 5)$. S_d contains one other, untrue state, s_d^1 with payoffs $(x_d^1, y_d^1) = (10, 10)$. If the DM reports s_d^0, he tells the truth and will obtain a payoff of $(5, 5)$ for himself and his partner. If, instead, he reports s_d^1, he tells a lie and obtains payoffs $(10, 10)$. Such a lie is called a *mutually beneficial lie (MBL)* as both affected individuals obtain a monetary benefit from the lie relative to the truth. Now imagine that S_d contains a further, untrue, state s_d^2 with payoffs $(x_d^2, y_d^2) = (15, 0)$. If the DM were to decide to report s_d^2, he would tell an *egoistic lie (EL)* as telling the lie increases his payoff but decreases the partner's payoff relative to the truth. If S_d also contains untrue state s_d^3 with payoffs $(x_d^3, y_d^3) = (0, 15)$ and the DM decides to report s_d^3, then he would tell an *altruistic lie (AL)* as the DM's payoff decreases but the partner's payoff increases relative to reporting the truth. Following this logic, there are nine lie types, as defined in Table 1.1.

Table 1.1: Lie Types

	$y_d^m > y_d^0$	$y_d^m = y_d^0$	$y_d^m < y_d^0$
$x_d^m > x_d^0$	Mutually beneficial (MBL)	Self-serving (SSL)	Egoistic (EL)
$x_d^m = x_d^0$	Weakly Altruistic (WAL)	Neutral (NL)	Harmful (HL)
$x_d^m < x_d^0$	Altruistic (AL)	Self-harming (SHL)	Mutually harmful (MHL)

Lie types are defined based on the payoff consequences of the lie relative to the truth for the decision-maker and the partner. A report is classified as a lie whenever $r_d \neq s_d^0$. x_d^0 stands for the decision-maker's payoff from reporting the true state s_d^0, x_d^m for the decision-maker's payoff from reporting an untrue state s_d^m. y_d^0 stands for the partner's payoff when the true state s_d^0 is reported and y_d^m for the partner's payoff when an untrue state s_d^m is reported.

I now define the DM's preferences. I allow for heterogeneity of lying preferences and represent this heterogeneity by a vector θ with

N entries $\theta_n \in \mathbb{R}$. The elements of θ, $\theta_1, \ldots, \theta_N$, refer to the different dimensions that enter the decision-making process such as the mental cost of lying, the strength of other-regarding preferences, or the effect of lie types. The vector θ thus defines those dimensions of the decision-making process that result in differences in lying preferences. The DM's preferences are described by the following utility function, with $m = 0, \ldots, M$:

$$u(s_d^m; s_d^0, \theta) := v(x_d^m, y_d^m, LT(x_d^m - x_d^0, y_d^m - y_d^0); \theta) - \mathbb{1}_{s_d^m \neq s_d^0} c_\theta$$

Component c_θ captures the psychological cost of lying, assumed to be constant for an individual but varying across individuals. Function $v(x_d^m, y_d^m, LT(x_d^m - x_d^0, y_d^m - y_d^0); \theta)$ captures several aspects of lying: the utility from payoff consequences for both the DM and the partner but also the effect of a lie type on utility. This last aspect is captured by $LT(x_d^m - x_d^0, y_d^m - y_d^0)$. Notice that both $LT(\cdot)$ and $v(\cdot)$ depend entirely on the monetary payoffs of a reported state and on the typology of lies, which is also consequence based. $LT(\cdot)$ can have both a positive or a negative effect on utility so that this can capture both utility as well as disutility across lie types. Importantly, utility itself is consequence based. The state that is publicly reported by the DM, r_d, is the state which maximises the DM's utility, taking into consideration whether reporting constitutes the truth or one of the nine lie types:

$$r_d \in \arg\max_{s_d^m \in S_d} u(s_d^m; s_d^0, \theta)$$

If S_d has only two elements, the choice is binary between telling a lie and telling the truth. If there are more than two states, the DM can decide between telling the truth (reporting s_d^0) or between telling one of several possible lies.

The framework's purpose is to guide the experimental design and the interpretation of the experimental results. As such, some identifying assumptions are needed in order to interpret observed behaviour effectively. The first identifying assumption (or property) is a tie

breaking rule. It states that if the utilities from reporting the true state s_d^0 is larger or equal to the utilities of reporting untrue states s_d^m, then the true state is reported. Notice that when the utilities are equal to each other, the tie-breaking property guarantees that the true state will be reported (the case when they are unequal is already governed by the the fact that r_d is the arg max).

PROPERTY 1 TIE-BREAKING RULE
For all $d \in \{1, \ldots, D\}$, if, for all $s_d^m \neq s_d^0$, $u(s_d^0; s_d^0, \theta) \geq u(s_d^m; s_d^0, \theta)$, then $r_d = s_d^0$.

The second identifying property states that the truth is weakly preferred to lying when there are no monetary consequences from lying relative to telling the truth.

PROPERTY 2 TRUTH IS WEAKLY PREFERRED
For all $d \in \{1, \ldots, D\}$ such that $x_d^0 = x_d^m$ and $y_d^0 = y_d^m, m \neq 0$, it holds that $u(s_d^0; s_d^0, \theta) \geq u(s_d^m; s_d^0, \theta)$.

For Property 2, when $u(s_d^0; s_d^0, \theta) = u(s_d^m; s_d^0, \theta)$ and the utilities are larger than those of any other state, Property 1 ensures that the true state is reported. Note that Property 1 governs situations with equality of utilities while Property 2 governs situations with equality of monetary payoffs. Together, these properties ensure that when a DM in the experiment reports an untrue state, reporting the untrue state must have generated strictly greater utility than reporting the truth. Observing an untrue state being reported thus reveals a strict preference for telling a lie.

Property 3 defines the heterogeneity of lying preferences in this framework. To ease notation, the difference in payoffs from lying and telling the truth will be given by: $\Delta x_d^m = x_d^m - x_d^0$ and $\Delta y_d^m = y_d^m - y_d^0$. They will be referred to as "relative payoffs of lying".

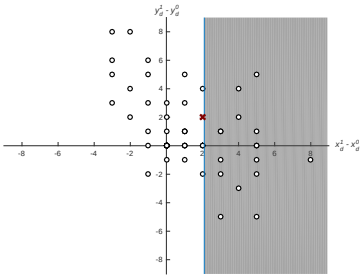
Property 3 Regions of inference
For all decision problems $d, \bar{d} \in \{1, \ldots, D\}$ such that $r_d = s_d^m$ with states $s_d^m \neq s_d^0$ and $s_{\bar{d}}^{\bar{m}} \neq s_{\bar{d}}^0$, at least one of the following has to hold:
(1) if $\Delta x_{\bar{d}}^{\bar{m}} \geq \Delta x_d^m$ then $u(s_{\bar{d}}^{\bar{m}}; s_{\bar{d}}^0, \theta) > u(s_{\bar{d}}^0; s_{\bar{d}}^0, \theta)$.
(2) if $\Delta y_{\bar{d}}^{\bar{m}} \geq \Delta y_d^m$ then $u(s_{\bar{d}}^{\bar{m}}; s_{\bar{d}}^0, \theta) > u(s_{\bar{d}}^0; s_{\bar{d}}^0, \theta)$.
(3) There exists $\alpha_\theta \in (0, 1)$ such that if $\alpha_\theta \Delta x_{\bar{d}}^{\bar{m}} + (1 - \alpha_\theta)\Delta y_{\bar{d}}^{\bar{m}} \geq \alpha_\theta \Delta x_d^m + (1 - \alpha_\theta)\Delta y_d^m$ then $u(s_{\bar{d}}^{\bar{m}}; s_{\bar{d}}^0, \theta) > u(s_{\bar{d}}^0; s_{\bar{d}}^0, \theta)$.

Property 3 states that if a DM preferred to lie for a particular decision-problem, he will also lie for decision-problems in which the relative payoff(s) of lying are weakly larger relative to those of the decision-problem where the DM already lied. Thus, increasing the gains of lying guarantees that there is no reversal in his preferences once the DM starts to lie. This implies that there exists a well defined "lie region" and that once a DM has moved to the "lie region", they will remain in that region.

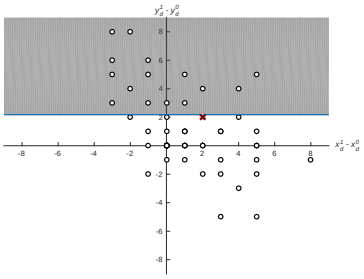
Depending on which of the sub-properties hold, "the gain" can refer to the DM's, the partner's or both relative payoffs. Property 3(1) implicitly assumes that the DM does not care about the partner's payoff as any increase in his own payoff realtive to the truth and the lie that he already told will lead him to continue to lie, irrespective of the change in the partner's payoff. Property 3(2) makes the opposite assumption in that the decision to lie is invariant to his own payoff but he will lie whenever the partner's payoff relative to the truth increases relative to the lie that he already told. Property 3(3) says that both changes in both payoffs matter. If the weighted sum of the payoffs relative to the truth payoffs is larger than the weighted sum of the relative payoffs of the lie that he already told, he will continue to lie. The parameter α_θ ensures that the payoffs can be weighted differently across individuals such that some DMs are allowed to care more about their own payoff than about the partner's payoff or vice versa. The reason why α_θ cannot take the extreme values of 0 and 1 is the following: if it were to take one of these values, it would subsume sub-properties (2) and (1) respectively.

Figure 1.1: Visualisation of Property 3

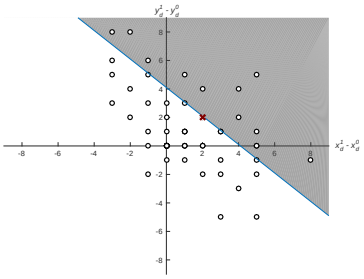
(a) Property 3(1)



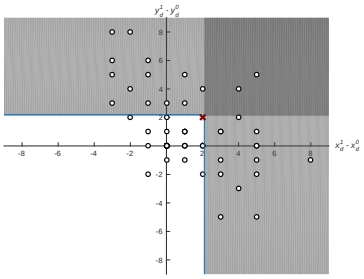
(b) Property 3(2)



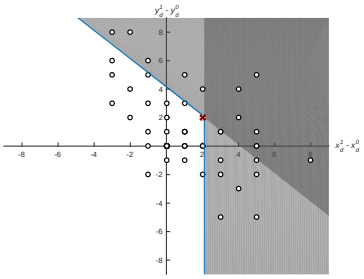
(c) Property 3(3), $\alpha_\theta = 0.5$



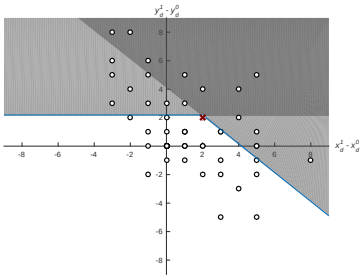
(d) Properties 3(1) and (2)



(e) Properties 3(1) and (3), $\alpha_\theta = 0.5$



(f) Properties 3(2) and (3), $\alpha_\theta = 0.5$



Implications of Property 3 visualised using an example. The red cross visualises the assumption that the DM lied for the payoff combination of $\Delta x_d = 2$, $\Delta y_d = 2$ where $\Delta x_d = x_d^1 - x_d^0$ and $\Delta y_d = y_d^1 - y_d^0$ are displayed on the horizontal and the vertical axis respectively. The grey areas indicate payoff combinations for which the DM has to lie, given Property 3. Darker grey tones for figures that display combinations of sub-properties indicate that sub-properties coincide with their prediction for this area.

In such a case, we could simply use only sub-property (3) to describe all types. However, if that were the case, we could no longer have combinations of the sub-properties (see paragraph below). Thus, we need all three sub-properties to be separate from each other and therefore do not allow α_θ to be 0 or 1.

It is important to emphasise that while Property 3 only requires that one of the sub-properties (1) - (3) has to hold, the sub-properties can also hold simultaneously. In such a case, the implicit assumptions are valid only for parts of the payoff space. To illustrate, imagine that Property 3(1) and 3(3) hold simultaneously. Then the DM should lie whenever $\Delta x_{\bar{d}} > \Delta x_d$ and whenever $\alpha_\theta \Delta x_{\bar{d}}^{\bar{m}} + (1-\alpha_\theta)\Delta y_{\bar{d}}^{\bar{m}} > \alpha_\theta \Delta x_d^m + (1-\alpha_\theta)\Delta y_d^m$. There is no contradiction for those cases where one of these holds but not the other, as the property does not make any statements about behaviour when the conditions, e.g. $\Delta x_{\bar{d}}^{\bar{m}} > \Delta x_d^m$, do not hold. Due to the two sub-properties holding for different payoff combinations, the union of the lie regions will contain a kink at the intersection of the two lie regions described by the two sub-properties which permits that behaviour can vary across lie types (see Figure 1.1(e)). When Properties 3(1) - (3) hold simultaneously, the lie region either coincides with that for the case when sub-properties (1) and (2) hold simultaneously or it will exhibit a kink at each of the two intersections.

Figure 1.1 illustrates the implications of Property 3 for an example. In the figure, the x-axis displays the payoff gain from reporting an untrue state s_d^1 relative to the true state s_d^0 for the DM, $x_d^1 - x_d^0$. The y-axis displays the equivalent for the partner, $y_d^1 - y_d^0$. This means that effectively, the payoffs from telling the truth are normalised to $(0, 0)$. For that reason, every dot represents a binary choice problem: it displays the monetary payoffs of lying relative to telling the truth, the normalised payoffs from telling the truth as well as the decision to lie. An empty circle indicates that we have not observed any choices for that decision problem. A red cross indicates that we have observed that a DM reported the lie for that choice problem. In the example, we have observed that the decision-maker reported a lie when the

relative payoffs were $\Delta x_d = 2$ and $\Delta y_d = 2$. Each sub-figure displays the lie region, indicated by the grey area, depending on which sub-property is assumed to hold. Recall that the lie region gives the set of decision-problems for which we anticipate that the DM will lie based on having observed that he lied for decision-problem d. Sub-figures (d) - (f) show the implications when more than one sub-property hold simultaneously.

The lie regions themselves are a valuable tool to predict behaviour based on having observed a DM's choices for a limited number of decision-problems. The tool can easily visualise how a DM is expected to behave for decision-problems that were not observed but that fall within the lie region. The lie regions thus provide the researcher with a tool to help anticipate behaviour without requiring additional data. At the same time, it can be used to falsify the model.

1.3 Experimental Design and Logistics

The experiment was designed with three requirements in mind. The first was to ensure that lying decisions are observable at the individual level without confounding beliefs in order to elicit lying preferences and assess the role of different types of lies. The second was the identification of types of decision-makers and the distribution of types in the data. Third, the design had to permit the clean separation of lying preferences from social preferences.

This section discusses first the logistics and then the design of each element of the experiment in detail.

1.3.1 Logistics

The experiment was conducted online with subjects recruited via the experimental platform Prolific.[1] In total, 103 subjects (51.5%

[1]Peer et al. (2017) and Palan and Schitter (2018) document the high quality of this platform relative to alternatives, both with respect to the participants and

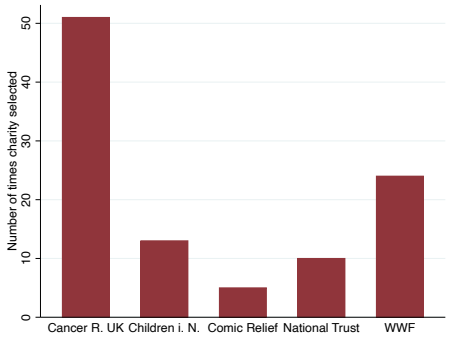
females, mean age was 36.6), which corresponds to roughly 88% of completed responses, passed the comprehension checks. Subjects spent on average 25 minutes on the experiment.

At the beginning of the experiment, subjects were informed that they would be matched with a charity that they could select from a list of five well-known and popular UK charities. The selected charity plays the role of "partner" throughout the experiment. The rationale for selecting a charity as partner is explained in detail in the subsection below. The majority of subjects who sign up to Prolific are from the the UK which is why UK charities were selected. While these charities are very popular in the UK, it is unlikely that someone with no ties to the UK is aware of their existence. For that reason, the subject pool was restricted to UK nationals. The five charities were: Cancer Research UK, Children in Need, Comic Relief, National Trust and World Wildlife Fund. These charities are active in the following areas respectively: medical research, child welfare, poverty relief, cultural heritage and wildlife support. The broad range of charities was selected to increase the probability that subjects cared about at least one of these charities. The charities were paid in the form of donations. Figure 1.2 shows how many subjects chose each of the charities. After subjects had selected their most preferred charity, they played the main experiment.

The experiment consisted of two main stages, a lying and a social preference game. Subjects were randomly allocated to starting with one of the games and then played the remaining game. After having completed the main stages, they were then asked to respond to a questionnaire that included demographic questions, a cognitive reflection test (CRT) and a Big 5 test. The CRT test contained four questions; three questions from a recent version of the test (Thomson and Oppenheimer (2016)) and one from the traditional CRT test (Frederick (2005)). The majority of questions was chosen from the recent version in order to reduce the likelihood that subjects already knew the answers by heart. This may be the case for questions from

the functionality of the platform.

Figure 1.2: Choice of Charities



The figure shows the number of subjects that chose each charity. The charities' names in full are: Cancer Research UK, Children in Need, Comic Relief, National Trust and World Wildlife Fund.

the traditional CRT test as subjects often encounter the traditional test in online experiments and may thus be familiar with the correct responses.[2] The question from the traditional CRT test was included to ensure comparability. The Big 5 test used was a short, ten item version (Rammstedt and John (2007)) in order to reduce the time spent on the experiment.

In addition to a show-up fee, subjects were paid based on their choices in two rounds of the experiment, one from each of the main stages. The random selection of one round for payment from each stage should reduce the likelihood of balancing behaviour with respect to payments to the charities. The rounds were selected by a random number generator. On average, subjects were paid 7.61 GBP.

[2]At the time the experiment was conducted, answers to the recent version of the CRT test were difficult to find online, thus reducing the chance of cheating.

1.3.2 Design

1.3.2.1 Lying game

Subjects played 60 rounds of a lying game developed for this paper, after having completed a game specific comprehension check. This game will be referred to as "the lying game". The lying game is informed by the framework introduced in Section 1.2. As such, each subject takes the place of a decision-maker who is matched with a passive partner, the charity, and each round forms one decision-problem d. The terms "round" and "decision-problem" will be used interchangeably from here on. In each round, a true state is privately observed and the DM is asked to report this true state but has the opportunity to lie.

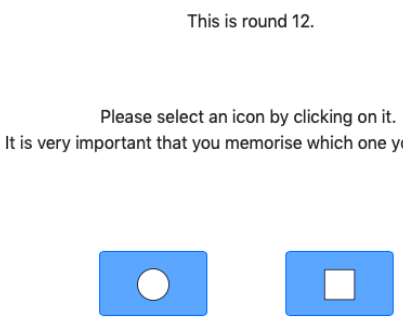
In the framework, subjects' choices affect a partner, the charity, who is passive.[3] Using a charity rather than another player as partner has several advantages in creating such a setting. First, the charity does not participate in the experiment and it is thus obvious to the subjects that their partner is passive which ensures that there are no first or second order beliefs about the partner's behaviour. Second, subjects actively select the charity which ensures that the partner is salient. Third, choosing from a list of several charities with different purposes ensures that they care, at least to some extent, about this partner. For control, I ask subjects to self-report how much they care about the charity that they have selected and charitable giving in general (82.52% reported that they liked the charity that they had chosen either moderately, very or extremely).

Each of the rounds consisted of two screens that the subjects saw sequentially. On the first screen, geometric shapes were displayed on buttons (for an example see Figure 1.3). These shapes consisted of easily recognisable geometric shapes, such as circles, squares, triangles, pentagons, diamonds and hexagons. Subjects were asked to choose

[3]Subjects selected the charity before they knew the specifics of the games or that it was a lying game.

one of the shapes by clicking on it and were told to memorise that geometric shape.

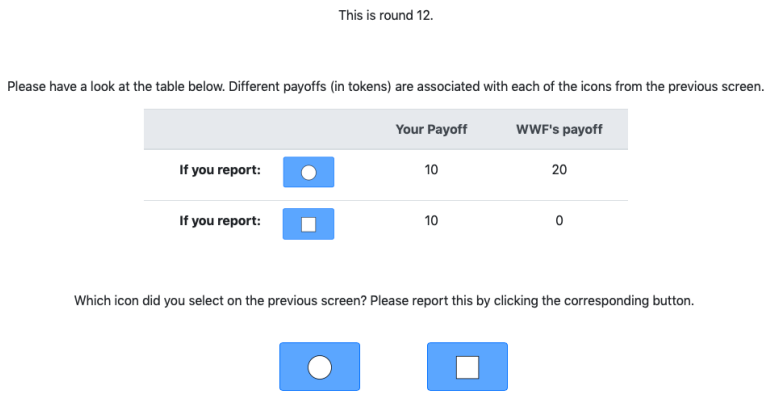
Figure 1.3: Example of the 1st screen of one of the rounds of the lying game



On the second screen, subjects saw the same shapes on buttons but now they were associated with monetary payoffs. Specifically, a table displayed the payoffs associated with each of the geometric shapes from the first screen (see Figure 1.4). Each shape thus corresponded to one state s_d^m in the framework. The payoff table showed one payoff for the subject, x_d^m, and one for the charity, y_d^m, for each of the shapes. Payoffs were given as an experimental currency (EC), subjects were paid in GBPs and were told the conversion rate of 5 tokens = 1 GBP in the instructions. In order to increase the salience of the partner, the name of the charity that the subject had chosen was displayed in the table.

At the bottom of the screen, subjects were asked to report which geometric shape they had chosen on the first screen. The choice on the first screen thus determined the true state s_d^0. If the subject reported the same shape on the second screen as he had clicked on in the first, his report was classified as telling the truth. If, however, he reported another state, the report was classified as a lie. As the experiment was conducted online and the initial choice as well the report were registered by the software, individual lying behaviour was observable

Figure 1.4: Example of the 2nd screen of one of the rounds of the lying game



to the researcher. Importantly, in conjunction with the use of the charity as partner, this implies that choices are observable at the individual level but do not suffer from (higher order) beliefs stemming from the expected reaction of the partner to the report. The type of the potential lie could be varied across rounds by changing the payoffs relative to the selected shape i.e. true state.

To illustrate, imagine that a subject is currently in round 12 of the lying game, $d = 12$. On the first screen of round 12, the subject has the choice between choosing a circle or a square (as in Figure 1.3). Further, imagine that he chooses the square. The true state of the twelfth round s^0_{12} is then "square" and there exists one untrue state s^1_{12} which is "circle". On the second screen, he is informed that the square is associated with 10 ECs for himself and 0 ECs for the charity while the circle is associated with 10 ECs for himself and 20 ECs for the charity (see Figure 1.4). Reporting the circle would constitute telling a *weakly altruistic lie (WAL)* and reporting the square would constitute reporting the true state, s^0_d. If, instead, the payoffs associated with the circle had been 20 ECs for both, reporting the circle would have constituted a *mutually beneficial lie (MBL)*.

The geometric shapes themselves were randomised across rounds and subjects to prevent shape-based effects. The order of the 60 rounds was randomised across subjects as well, to reduce order effects. To prevent that subjects simply clicked through the rounds rather than making choices based on their lying preferences, buttons changed positions between rounds and between screens within rounds so that it was impossible for someone to keep the mouse cursor in the same position and then click through the whole experiment.

All subjects played the same 60 rounds. Rounds differed with respect to the particular payoffs and the number of available choices. In 50 rounds, DMs were faced with a binary choice between telling the truth, reporting s_d^0, and telling a lie, reporting s_d^1. In ten rounds, more than one lie type was available for reporting in order to be able to directly research preferences between lie types. Here, DMs were given a choice between telling the truth and telling one of multiple lies rather than a binary choice between telling the truth and telling a lie. The specification of each round are given in Table 1.5 in the Appendix.

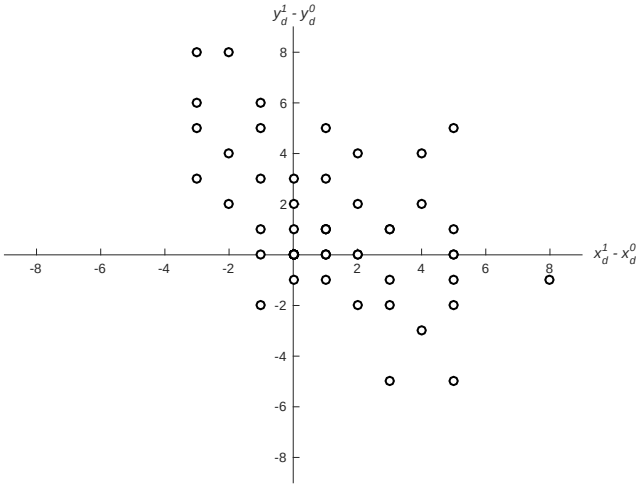
The framework allows for heterogeneity in the utility function and one of the goals of the experiment is to identify whether and where such heterogeneity exists. Therefore, the payoffs across rounds were systematically selected to maximise variation in behaviour across the lie type space. Several pilots examined behaviour for a wide range of payoffs. To tease out the heterogeneity, more rounds were located close to payoff combinations where the differences across subjects were expected to become apparent, based on results from the pilots, and where a fine payoff grid was thus necessary. Figure 1.5 shows the space of payoffs for rounds with two states, s_d^0 and s_d^1. The x-axis displays the change in the DM's payoff from reporting s_d^1 compared to reporting the true state s_d^0, and the y-axis displays the same for the charity's payoff. Each dot represents at least one round. To control for possible level and inequality effects, some payoff differences were used multiple times so that multiple rounds, with different absolute

consequences but identical relative ones, can be represented by one dot. For example, imagine that in one round d with two states, the true state's payoffs are $(x_d^0, y_d^0) = (5, 5)$ and the untrue state's payoffs are $(x_d^1, y_d^1) = (3, 7)$, while in another round $\bar{d}$ the payoffs are $(x_{\bar{d}}^0, y_{\bar{d}}^0) = (7, 3)$ and $(x_{\bar{d}}^1, y_{\bar{d}}^1) = (5, 5)$. In both rounds, the payoff differences are $x_d^1 - x_d^0 = x_{\bar{d}}^1 - x_{\bar{d}}^0 = -2$ and $y_d^1 - y_d^0 = y_{\bar{d}}^1 - y_{\bar{d}}^0 = 2$. Both rounds are therefore visualised by the same dot in Figure 1.5. However, in the second round, lying generates equality of payoffs while in the first round, lying generates inequality. These inequality concerns are likely to affect lying behaviour. To control for such considerations, the payoffs for the rounds have been selected to ensure that each lie type is represented by situations that vary with respect to whether a potential lie reduces or increases inequality.

Behaviour in the experiment can be treated as revealed preferences when the framework is applied to the experiment, as explained in Section 1.2. Specifically, the identifying assumptions, Properties 1 and 2, imply that any untrue state that is reported reveals a strict preference for lying in that round and thus allow to map behaviour into preferences. It is then possible to test whether there exist subjects who always choose to report the true state s_d^0 and who can be classified as never-liars and whether some subjects lie occasionally. The reported states and the attached payoffs can be used to examine whether there exists a payoff combination for which a subject switches from telling the truth to telling a lie. Based on these inflection points, subjects can be classified into several types of DMs, θ. Finally, behaviour in the experiment can be assessed on whether changes between lying and telling the truth are systematic and if yes, whether the regions of inference property, Property 3, can describe behaviour.

The design introduced above was selected with the aim to isolate lying preferences. A possible concern with the lying game could be that when subjects misreport which icon they had selected, this might be due to memory issues. However, subjects had to memorise only one geometric shape per round for a few seconds and the shapes were easy to remember. Systematic and frequent choice errors are

Figure 1.5: Payoff space for binary questions



All rounds with two states (true state s_d^0 and an untrue state s_d^1). The horizontal axis displays $x_d^1 - x_d^0$ and the vertical axis displays $y_d^1 - y_d^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Some rounds use the same payoff differences but have different levels of payoffs so that some dots describe multiple rounds.

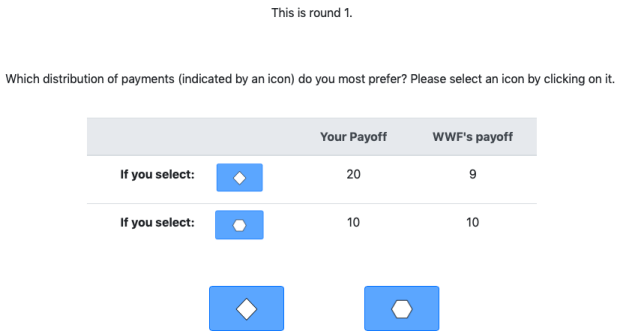
thus unlikely. In addition, more than a third of subjects did not lie at all which suggests that a large fraction of subjects was able to correctly remember the choice they had made on screen one in each of the rounds. This supports the claim that it was easy to remember the initial icon choice. Moreover, a large majority of subjects was very systematic in their behaviour, reducing the likelihood of memory errors as primary driver and further illustrating the success of the design in capturing behaviour across lie types.

1.3.2.2 Social preference game

When examining lying behaviour in settings with two individuals, a key point of interest is whether subjects' responses in the lying game perfectly, or highly, correlate with their social preferences and are thus not so much driven by lying considerations but by their social preferences. For this reason, the experiment contained a social preference game which has been designed to address this concern directly.

The social preference game was thus designed to be directly comparable to the lying game. Consequently, this game also consisted of 60 rounds. Each round was identical to the corresponding round from the lying game with respect to the choice sets. However, to elicit social preferences, there was one exception: subjects were only presented with the second screen which displayed the payoff table. Here, they were asked to select their most preferred option. The key difference is that because subjects only see the second screen, there is no truth benchmark for the round any more. Subjects are free to choose their most preferred payoff bundle without any concerns about lying or truth-telling entering the decision-making process. This allows the researcher to directly contrast choices in a distributive setting to those in a lying setting. Figure 1.6 displays an example screen from one round of the social preference game.

Figure 1.6: Example screen of one of the rounds of the social preference game



1.4 Results of the Lying Game

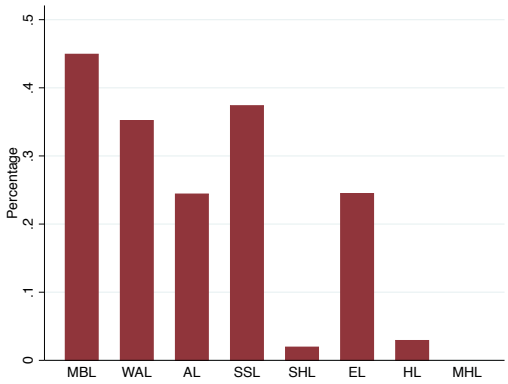
This section presents the results of the lying game. It starts by presenting the aggregate experimental results to facilitate comparison with the literature. It then examines whether lie types matter, measures the heterogeneity of decision-makers and the distribution thereof and examines whether the properties hold in the data. The relationship between social preferences and lying preferences in this experiment is discussed in the section that follows.

1.4.1 Aggregate results

In the experiment, 31.11% of subjects lied on average per question. This percentage varies drastically across lie types, demonstrating the importance of subdividing lies into types. Figure 1.7 shows the mean percentage of lies told across the lie types. The percentage of lies averaged by lie type is given by the red bars. They show that the highest percentages of lies, 35% to 45%, occurred for lie types where at least one out of the two subjects benefits from the lie without the other suffering, namely *mutually beneficial*, *weakly altruistic* and *self-serving lies*. *Altruistic* and *egoistic lies* have similar rates of lying

of around 25%. Very few lies were told that harm either the DM or the charity relative to the payoff connected to the truth, i.e. *self-harming*, *harmful* or *mutually harmful lies*. This would have been different if another context had been provided, for example in-group out-group effects, where it would have been possible that a DM would have been willing to incur a small loss to severely harm someone from an out-group. Differences in behaviour across chosen charities are shown in the Appendix.

Figure 1.7: Percentage of liars by lie type



The percentage of liars averaged by lie type is given by each bar. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *self-serving lie* (SSL), *self-harming lie* (SHL), *egoistic lie* (EL), *harmful lie* (HL) and *mutually harmful lie* (MHL).

Next, I compare the percentages of lies told in the literature to those found in the lying game. In order to be able to compare the results across lie types, my results need to be compared with several papers as most of them examine a small number of lie types, often one, per paper. The results are displayed in Table 1.2. As it is a seminal paper, the first column shows the results of Gneezy (2005). I also report the results of Hurkens and Kartik (2009) as they supplement

Gneezy (2005)'s findings by increasing the variation of lies considered in the paper. To obtain a sense of percentages of lies told for lies that are not egoistic, I also include Biziou-van Pol et al. (2015) and Erat and Gneezy (2012).

The table reveals that there is substantial variation in the percentage of lies, between as well as within lie types. For example, to see the variation between lie types, Biziou-van Pol et al. (2015) find that 83% of subjects told a *mutually beneficial lie* compared to 23% who told an *altruistic lie*. To see the variation within lie types across types, compare Erat and Gneezy (2012)'s finding that between 49% and 65% of subjects told a MBL with the 83% from Biziou-van Pol et al. (2015). The variation suggests that the percentage of lies depends both on the specific payoffs linked to a question and on the lie type. It is thus important to include many specifications of a lie type that differ with respect to the relative payoffs, equality concerns, level effects etc., for each lie type in an experiment. Note that the ranges of percentages of lies told by lie type are compatible with the results of the papers listed. For an analysis of the relationship between percentages of lies told and covariates such as cognitive reflection test or Big 5 scores, see the Appendix.

1.4.2 Do lie types matter empirically?

The previous section has provided evidence that people behave differently across lie types. This section formally analyses the degree to which lie types matter and how they should be included in lying frameworks.

To this end, I examine the degree to which two baseline utility models from the literature can explain the experimental data as a whole. I then compare their performance to two models that differ with respect to the role of the lie type in the utility. The first model allows the constant cost of lying, c_θ, to vary across lie types while the second model additionally allows the lie type to affect the utility from the payoffs of a lie i.e. where the lie type enters $v(\cdot)$. Comparing

Table 1.2: Percentage of liars by lie type: Comparison with literature

Lie Type	Gneezy (2005)	Hurkens & Kartik (2009)	Biziou-van-Pol et al. (2013)	Erat & Gneezy (2012)	This paper
EL	17 - 36%	38 - 47%	NA	37%	12 - 40%
AL	NA	NA	23%	33%	4 - 35%
MBL	NA	NA	83%	49 - 65%	40 - 50%
SSL	NA	NA	NA	52%	32 - 41%
WAL	NA	NA	NA	NA	32 - 42%
SHL	NA	NA	NA	NA	2%
HL	NA	NA	NA	NA	3%
MHL	NA	NA	NA	NA	0%

Example percentages of lies told per lie type in the literature compared to this paper. The acronyms stand for *egoistic lie* (EL), *altruistic lie* (AL), *mutually beneficial lie* (MBL), *self-serving lie* (SSL), *weakly altruistic lie* (WAL), *self-harming lie* (SHL), *harmful lie* (HL) and *mutually harmful lie* (MHL). NA signifies that a paper did not include the lie type listed in the row. When a paper examined at least two instances of a lie type, the range of percentages of lies told by subjects is given.

the baseline models to the augmented models can tell us whether lie types have a significant impact on preferences while comparing the augmented models with each other can tell us how lie types enter preferences.

The baseline model has been selected to fit with the most basic theory of lying: when deciding whether to report the true state s^0 or an untrue state $s^m, m \neq 0$, the DM compares only his own payoff from lying to his payoff when telling the truth but faces a constant cost of lying.[4] The DM's utility from reporting the untrue state s^m, with the utility from telling the truth normalised to zero, is then given by the utility from the monetary payoff from lying compared to that from reporting the true state minus a constant cost: $u(s^m; s^0, \theta) = \beta_\theta[x^m - x^0] - \mathbb{1}_{s^m \neq s^0} c_\theta$. Papers such as Gibson et al. (2013) use models in this spirit as a baseline. As most papers

[4]For ease of readability, the d subscript denoting the decision problem has been omitted throughout this section and the next.

acknowledge that the DM might also care about a person other than the DM who is affected by the lie, I also use an extended baseline model. In the extended model, the DM's utility depends on both his own payoff as well as that of the partner and contains a constant cost: $u(s^m; s^0, \theta) = \beta_\theta[x^m - x^0] + \gamma_\theta[y^m - y^0] - \mathbb{1}_{s^m \neq s^0} c_\theta$. To report an untrue or the true state in such settings is a binary decision and is therefore modelled by logistic regressions. The baseline models' equations are given by:

Baseline model :

$$E[Y_\theta | x^0, x^m, \theta] = \frac{1}{1 + exp^{-(\beta_\theta(x^m - x^0) + \tilde{c}_\theta)}} \tag{1.1}$$

Extended baseline model :

$$E[Y_\theta | x^0, x^m, y^0, y^m, \theta] = \frac{1}{1 + exp^{-(\beta_\theta(x^m - x^0) + \gamma_\theta(y^m - y^0) + \tilde{c}_\theta)}} \tag{1.2}$$

where $Y_\theta = 1$ if a DM of type θ lied, $c_\theta = -\tilde{c}_\theta$ is a constant cost of lying and β_θ and γ_θ are weights on the monetary payoffs.

To analyse how well the models can explain the data, I perform an in-sample prediction exercise where the models were fitted to the data using logistic regressions. For the prediction, I use the whole sample to fit the data. The performance of the models is evaluated based on the size of the mean squared forecast error (MSE).

To assess the performance of the model with varying constant costs, I add dummies for the lie types to the extended benchmark model. This reduces the MSE to 0.1969 (min SE of = 0.0004 and max SE = 0.9615). The improvement in the MSE, compared the benchmark and the extended benchmark models, is marginal. This suggests that lie types do not affect lying preferences via shifts in the constant cost of lying.

I then examine the explanatory power of the model that permits lie types to affect the utility from the payoffs, $v(\cdot)$. To capture this, I split the data by lie type and fit the model to each lie type separately.

In order to compare the the performance of the lie type specific model to that of the other models, MSEs are averaged across the lie types.

Table 1.3 displays the MSE of each model as well as the minimum and maximum squared error for both prediction exercises. It shows that the benchmark and the extended benchmark models have a similar performance to each other. If we simply used those models, this would suggest that the partner's payoff plays only a limited role for the decision to lie. When we allow for the constant cost to vary across lie types, there is a very slight improvement in the MSE which suggests that lie types do not impact lying preferences via the constant cost of lying. However, when we examine the MSE of the model that allows for variation in the utility of payoffs across lie types, there is an improvement of 21.5% in the MSE. These results show first, that lie types play a significant role for lying preferences and second that they enter these preferences by interacting with the utility from monetary payoffs rather than via shifts in constant costs.

1.4.3 Heterogeneity of preferences across individuals

My theoretical framework assumes that individual lying preferences are heterogeneous and the inference region property, Property 3, imposes structure on how this heterogeneity can look like. Specifically, the property suggests that there exist around six groups of decision-makers and lays down expected behaviour for these groups. This subsection examines whether the property can capture subjects' behaviour in the experiment.

To systematically and objectively examine subjects' heterogeneity, I employ an unsupervised machine learning algorithm which creates a partition of subjects into groups.[5] Specifically, I use a k-means clustering analysis which is a statistical pattern recognition tool that,

[5]While the framework provides a prior for the number of groups and their behaviour, this section will approach the issue objectively, i.e. without imposing these priors, to prevent that the priors bias results in favour of the framework.

Table 1.3: In-sample prediction results

	MSE	min SE	max SE
Baseline model	0.2108	0.0409	0.6366
Extended baseline model	0.2036	0.0146	0.7727
Varying constant cost model	0.1969	0.0004	0.9615
Varying utility from payoffs model	0.1590	0.0190	0.2470

In-sample prediction results for the four models considered. Performance of the models is given by the mean squared error (MSE) of the forecast as well minimum and maximum squared errors (min SE and max SE, respectively). Row "Baseline model" displays the performance of the model that only includes the decion-maker's payoff. Row "Extended baseline model" displays the performance of the model that also includes the partner's payoff. Row "Varying constant cost model" displays the performance of a model that includes the decision-maker's and the partner's payoff and allows the constant cost to vary across lie types. Row "Varying utility from payoffs model" displays the performance of a model that includes the decision-maker's and the partner's payoff and allows the utility of the payoffs to vary across lie types.

in economics, is commonly used in time-series econometrics (see for example Falat and Pancikova (2015), Bagnall et al. (2003), Focardi and Fabozzi (2001)). k-means is one of the most popular classification algorithms.[6]

[6]Many papers in economics use a related method called mixture models. Mixture models are a widely used tool in economic analysis, primarily to detect and model heterogeneity (see for example Cameron and Heckman (1998)), such as identifying utility function shapes across heterogeneous individuals. K-means "is closely related to the EM algorithm for estimating a certain Gaussian mixture model" (p. 510, Hastie et al. (2009)). Specifically, mixture models make a probabilistic assignment of observations to the groups while the k-means algorithm uses deterministic assignments. When the variance of the Gaussian density becomes zero, the two methods coincide (Hastie et al. (2009)). As I am interested in deterministic group assignments, k-means is the preferred method, here.

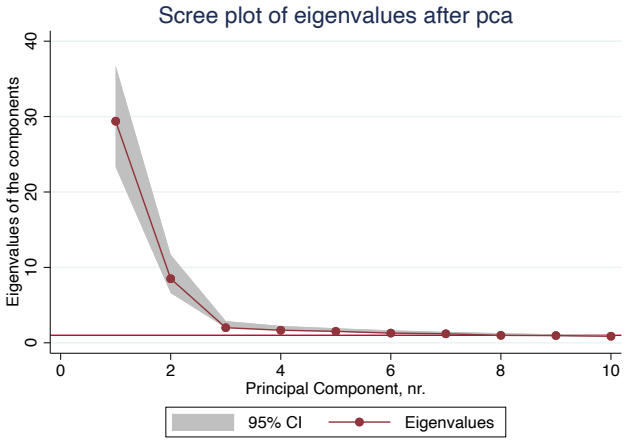
The algorithm works in the following way: First, the researcher specifies the number of clusters k and initialises the location of the centroids of the clusters randomly (a glossary of the terminology is provided in the Appendix). All subjects are then categorised as belonging to one of the clusters. A subject is allocated to the cluster with the centroid that is the closest to it, as measured by the Euclidean distance (Hastie et al. (2009)). The algorithm then determines new centroids of the clusters based on shifting the centroid to minimise the distance to all members that have been allocated to it. These steps are repeated until the clusters are "stable": updating the centroids does not affect group membership or the position of the centroids. The method is unsupervised in the sense that the true group membership is unknown to the researcher.

The algorithm is suitable to situations such as identifying heterogeneity of preferences for the following reason: the algorithm requires to know only the number of groups it is looking for and the variables on which it bases the group selection. Furthermore, if the number of heterogeneous groups is unknown, the method can be combined with information criteria with which the number of groups can be selected. The advantage of using a k-means cluster analysis over alternative methods, such as Bayesian methods, is thus that results do not depend on an informative prior beyond knowledge of the variable in which heterogeneity is expected.

In order to weight all variables evenly, variables are standardised by subtracting the mean and dividing the variables by their standard deviation. I then perform a principal components analysis (PCA) on the binary decision to lie of each of the 60 questions and retain only the main principal components, also called factors. PCA is conducted to prevent that clusters are biased towards variables that explain little of the data. Specifically, PCA allows the researcher to reduce the number of variables while preserving the informational content of the 60 lying decisions made by each subject. This procedure is also known as factor model analysis. Factor models are widely used in economics and finance (see for example Engle and Watson (1981),

Chamberlain (1983), Stock and Watson (2005)), typically to reduce the number of parameters that have to be estimated. To identify the number of components that should be retained, a scree plot of the eigenvalues after PCA is created. A scree plot is a figure which plots the eigenvalues of the principal components. One then looks for the so called "elbow point" where the information gain of adding another component levels off. Figure 1.8 shows that only the first three components have notable explanatory power and I therefore only include these three components in the k-means analysis.

Figure 1.8: Scree plot of the 10 largest components in lying game PCA analysis

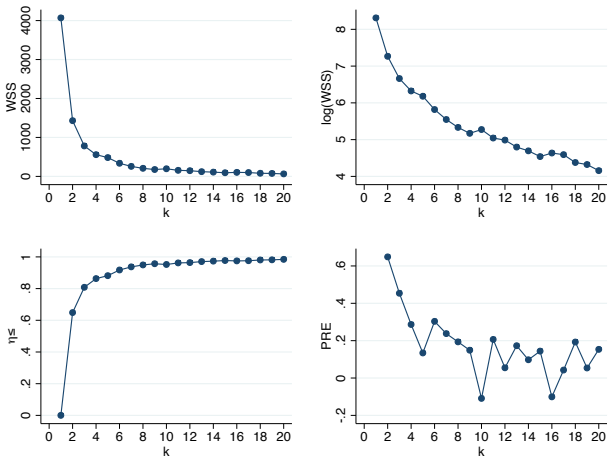


The x-axis shows the largest components (descending) and the y-axis the size of the eigenvalues.

Based on the regions of inference property of the framework, we expect a number of types, and thus clusters, close to six (see the possible sub-property combinations for Property 3). To objectively assess the number of types of DMs in the data, I conducted an initial analysis of how many groups are ideal for the analysis. This consists of repeating the k-means exercise for $k = 1, ..., N$ groups (here, N=20)

and compare the total sum of squares (TSS, calculated when $k = 1$) to the within sum of squares (WSS). The optimal number of groups can be found where the improvements in explanatory power are levelling off. Figure 1.9 shows the performance of the different numbers of clusters. The figure indicates that the number of clusters k should be set equal to six as the informational gains of adding another group level off at around $k = 6$.

Figure 1.9: Performance indicators by cluster number in lying game k-means analysis

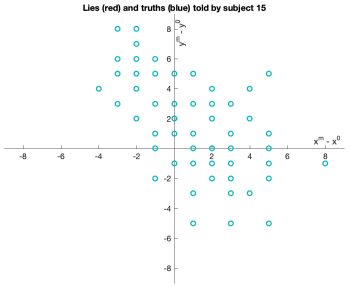


Plot of within sum of squares (WSS), log WSS, η^2 and proportional reduction of error (PRE) for number of clusters $k = \{1, ..., 20\}$. The figures show that gains from adding another cluster level off at $k = 6$.

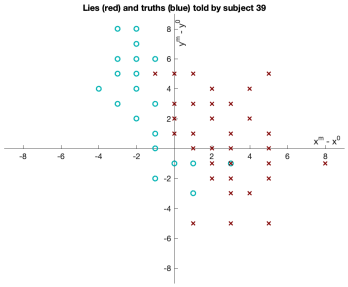
The clusters obtained via the k-means analysis with $k = 6$ can explain circa 91% of the variation in the data ($\eta^2 = 0.91$; η^2 is a goodness of fit analysis similar to R^2). We can thus say that behaviour across individuals is highly heterogeneous. The heterogeneous groups of DMs are represented graphically by Figure 1.10. It shows the

Figure 1.10: Representative DMs by behavioural lying cluster

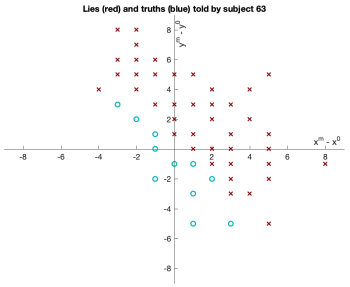
(a) Never-liar



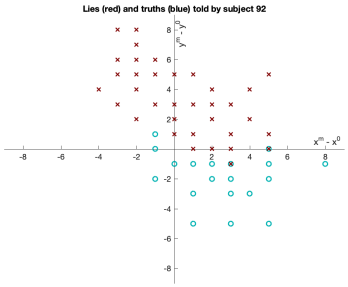
(b) Lies mainly when DM gains



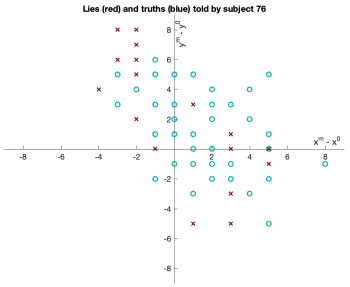
(c) Lies for both DM's and charity's gain



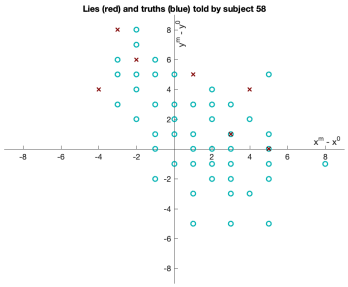
(d) Lies when the charity gains



(e) Balancing



(f) Non-systematic behaviour



Each panel of the figure shows a representative subject from each identified cluster in the lying game. The x-axis displays $x^m - x^0$ and the y-axis displays $y^m - y^0$, with $m = 1$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Each dot presents an available lie relative to the truth which is standardised to the origin (0,0). Red crosses indicate that the subject lied and a teal dot that they chose the truth instead.

binary choice between lying and telling the truth across all questions with a binary choice (one lie and one truth available). Each panel in the figure represents one of the six clusters.

The group easiest to identify, both visually and statistically, is that of never-liars (see Figure 1.10(a)). This group contains 44 out of the 103 subjects (some of them misreport a handful of times throughout the experiment but these appear to be errors rather than choices, see Section 1.8.4.2 in the Appendix). Of those, 28 truly never lied corresponding to 27% of the sample. This proportion of never-liars is exactly what we should expect based on the percentages of never-liars found in the literature (for example Erat and Gneezy (2012) find that 35% of subjects don't lie even when this would have resulted in a Pareto improvement in payoffs). The second largest group, with 15% of subjects, is that of subjects who lie whenever there own payoff (weakly) increases, regardless of the effect on the charity (see Figure 1.10(b)). The third largest group with 14% of subjects is that of subjects who lie both for the gain of the charity and the DM (see Figure 1.10(c)). Subjects in this group differ with respect to the degree with which they weight the DM's gains/losses relative to that of the charity, corresponding to differences in α_θ. In the fourth group, which also contains 14% of the subjects, are subjects who lie only when the charity gains from the lie (see Figure 1.10(d)). Subjects in this group show greater within-group variation than those in the previous three groups. Subjects in groups five and six are similar in so far that their behaviour is non-systematic (see Figure 1.10(e) & (f)). Together, they account for around 15% of all subjects (16 out of 103). Group 5 contains only two subjects where one of them said that they balanced between helping the charity and helping themselves. It is noticeable that on average, subjects in groups 5 and 6 made more mistakes in the comprehension check of the lying game than subjects from the other groups (43.75% in groups 5 and 6 made at least one mistake compared to 33.4% in the other groups). This may indicate that these subjects did not pay attention to the instructions, providing an explanation of the randomness of behaviour.

Having confirmed that there exist types of DMs who lie in some situations, we can now examine whether the regions of inference property, Property 3, holds for these types of DMs. The property states that once a DM has started to lie for a combination of payoffs defined by the states s^0 and s^m, he should continue to lie for every other decision problem in which a state $s^{\bar{m}}$ has a payoff bundle for which either $\Delta x^{\bar{m}}$ and/or $\Delta y^{\bar{m}}$ is larger than that of the state for which he lied, Δx^m and Δy^m, ceteris paribus. This ensures a kind of monotonicity in the space of Δx and Δy. To examine whether we observe the predicted behaviour in the experiment, we can examine Figure 1.10. Visually, we should observe red crosses, i.e. lies, only in the upper quadrants and in the lower right quadrant if Property 3 holds, with Property 2 as identifying assumption. To illustrate, imagine that someone lies for a mutually harmful lie. Then, Property 3 implies that he should also tell a neutral lie where the payoffs of the untrue state coincide with those of the true state (as this constitutes an increase in both Δx and Δy and as at least one of (a) - (c) has to hold). However, Property 2 states that for a neutral lie, the DM should always obtain more utility from reporting the true state. We thus have a contradiction. To ensure that both properties always hold, we need to have that DMs do not tell a mutually harmful lie and only start lying when either Δx and/or Δy is larger than zero. Inspecting the figure shows that this is the case for all representative agents across the identified groups. Behaviour in the experiment thus confirms the predicted behaviour from the theoretical framework, Property 3, when assuming that Property 2 holds as identifying assumption.

To show that the six individuals that have been presented in Figure 1.10 are representative of their clusters, I provide the same figures as above but instead of showing the decisions of an individual, they show what percentage of the subjects of the cluster lied for each question (see Figure 1.22 in Appendix 1.8.4.3).

Note that I find very systematic behaviour for four of the groups, which indicates that demand or experimenter effects did not play a

role for decisions. If the instructions or the set-up of the experiment had biased behaviour in a particular direction, we should observe that behaviour is biased in one direction. However, the largest group, never-liars, contains the expected number of subjects given the literature so that it does not seem that behaviour was biased in that direction. The other groups that show systematic behaviour are of nearly identical size (15, 15, and 14 subjects respectively). It is highly unlikely that the instructions primed behaviour to follow these three different patterns as well as leading to a representative number of never-liars. Thus, experimenter and demand effects should not be of particular concern.

Finally, I assess how meaningful the identified heterogeneity across subjects is. To this end I conduct two of sample prediction exercises and contrast their performance. For the first exercise, I ignore heterogeneity and for second, I account for it. I can then compare the accuracy of the forecasts and obtain a measure for how much of a difference including heterogeneity makes.

For both exercises, I re-conduct the k-means analysis but this time I only use 40 out of the 60 decision-problems to classify subjects. The 40 questions were selected at random across all lie types; results are robust to varying the 40 questions that are selected. For each subject, I then fit the extended baseline model to the data for the 40 selected questions from all subjects other than the subject of interest. I then predict the behaviour of the subject of interest for those 20 questions that were excluded from the model fitting exercise.

In the first prediction exercise, I do not account for heterogeneity in that I do not discriminate between decision-maker types when fitting the model.

I contrast the performance of the first prediction exercise with that of a prediction exercise which accounts for subject heterogeneity. For this second exercise, I re-conducted the out-of-sample prediction exercise. The difference to the first prediction exercise is that this time when fitting the model, the data for the 40 questions came from only

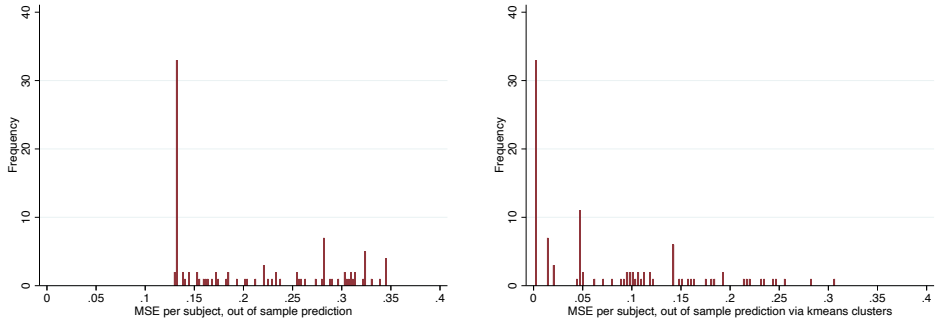
those subjects who were classified as belonging to the same cluster as the subject of interest.[7] As before, I then use the fitted model to predict the decisions of the subject of interest for those 20 questions that were not included in the model fitting and the k-means exercises.

Figure 1.11 reports the out-of-sample forecast performance of both prediction exercises, where prediction accuracy is given by the mean squared forecast error (MSE). Comparing the two panels, we can see that the variance of the MSEs as well as the mean size of the forecasting errors is reduced in the exercise that accounts for heterogeneity compared to the one that does not, panels (b) and (a) respectively.

Figure 1.11: Out-of-sample performance without and with heterogeneity in preferences taken into account

(a) Subject heterogeneity ignored

(b) Subject heterogeneity taken into account



Histogram of each subject's mean squared error (MSE) in pseudo out-of-sample prediction exercise. Panel (a) shows that the MSEs are larger when heterogeneity in preferences is ignored during the forecasting exercise compared to panel (b) where it is taken into account.

The improvement in the MSEs corresponds to a gain in forecasting

[7] As this is an out-of-sample forecast, the individual's behaviour itself is always dropped from the cluster before fitting the model to that group.

accuracy of more than 60 percent. To illustrate, the prediction exercise in panel (a) correctly predicted a subject's behaviour in 79.4% of decision-problems on average. The out-of-sample forecasts that did account for heterogeneity, panel (b), in contrast on average correctly predicted behaviour in 92% of the decision-problems. This constitutes a large improvement.

In order to examine whether the gains from adding heterogeneity to the out-of-sample prediction exercise are robust the number of clusters specified, I reran the k-means analysis with $k = 5$ and $k = 7$ clusters. The gains from accounting for heterogeneity are robust to slightly changing the number of clusters (see Figure 1.21 in the Appendix).

In sum, in this section we have learned that lie types enter lying preferences via the utility from payoffs, that subjects are highly heterogeneous and this heterogeneity is both systematic and consistent with that postulated by the framework. Finally, accounting for this heterogeneity in out-of-sample prediction exercises yields large and significant gains in the accuracy of the forecasted behaviour.

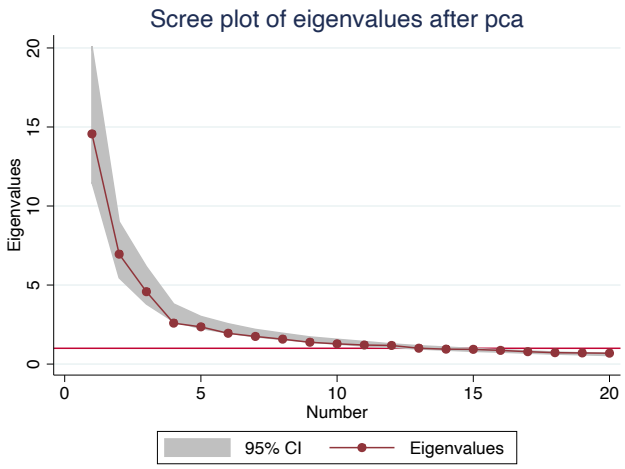
1.5 Lies and social preferences

In this section, I examine whether subjects' behaviour changed between the lying and the social preference games and if they did, whether patterns of behaviour changed systematically. The first subsection presents a k-means cluster analysis of behaviour in the social preference game where an algorithm allocates subjects to groups. The group composition is then compared against that of the groups defined by the lying game behaviour. Following the between group analysis, the second subsection analyses the interaction of social and lie preferences at the individual level.

1.5.1 Are behavioural groups in the lying game comparable to those in the social preference game?

Section 1.4 presented a k-means cluster analysis of lie behaviour. Here, I conduct the same analysis but using behaviour from the social preference game, only. Before conducting the k-means analysis, I conduct the same PCA exercise as before, where the 60 decisions are condensed into the principal components. Here, the four largest principal components are needed to capture the variation in the data (see Figure 1.12). The same seed is used for the initialisation of the clusters and as the aim of this section is to compare the groups from the lie and the social preference behavioural clusters, the number of clusters k is pre-specified to be equal to six.

Figure 1.12: Scree plot of 20 largest components in social preference game PCA analysis



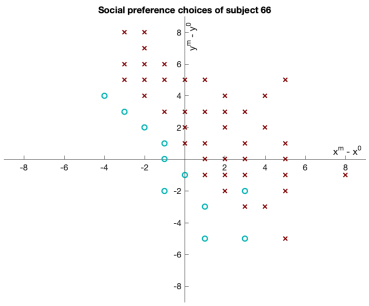
The x-axis shows the largest components (descending) and the y-axis the size of the eigenvalues.

Figure 1.13 shows the behaviour of representative individuals for each of the six groups. As with the equivalent figure in the section above, the figures show which alternatives the subject chose for those questions that entailed a binary choice. The origin symbolises the reference point's normalised payoffs, where the reference point is the option that is coded as the truth in the lying game. The x-axis shows the change in the DM's payoff relative to the reference point and the y-axis the corresponding values for the charity. For example, this means that where both values are positive, picking that alternative would lead to gains for both the DM and the charity relative to the reference point. Red crosses indicate the subject selected that option while teal dots indicate that the subject selected the reference point. I am referring to that option as the reference point only for ease of comparability. If social preferences were the only driver behind behaviour in the lying game, we should expect to the see similar pictures as for the lie groups, except for the never-liar group. If there exists a constant cost of lying, then the patterns of behaviour should be the same but the red crosses should be shifted downward (until the origin) in the social preference game. Of course, it is possible that the same behavioural groups exist but that different members populate them. This will be examined in Subsection 1.5.2.

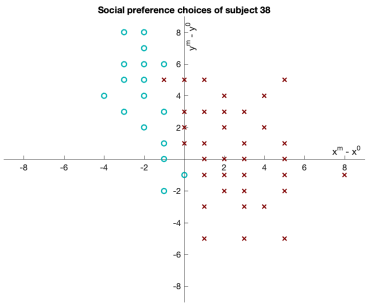
The largest cluster identified by the k-means algorithm contains subjects whose behaviour looks as if they were maximising payoffs of both the DM and the charity with equal, or at least close to equal, weights (see Figure 1.13(a)). This cluster contains nearly 47% of the subjects. In the second largest cluster, whose members constitute 20.4% of the sample, subjects choose alternatives to maximise the DM's payoff (Figure 1.13(b)). In the third cluster, which is the same size as the previous one, subjects behave to maximise the charity's payoff (Figure 1.13(c)). Subjects in cluster four, which correspond to 8.7% of the sample, seem to balance behaviour in so far that they sometimes maximise the charity's and sometimes the DM's payoff and there is no clear pattern (Figure 1.13(d)).

Figure 1.13: Representative DMs by behavioural social preference cluster

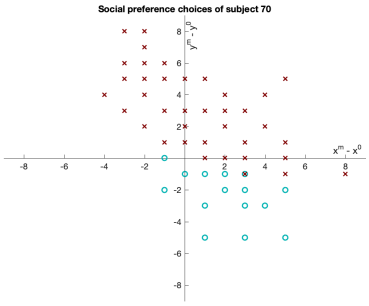
(a) Picks alternatives that maximise DM's and charity's combined payoffs



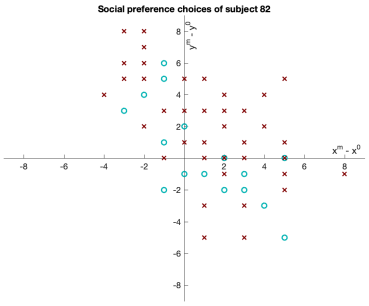
(b) Picks alternatives that maximise DM's payoffs



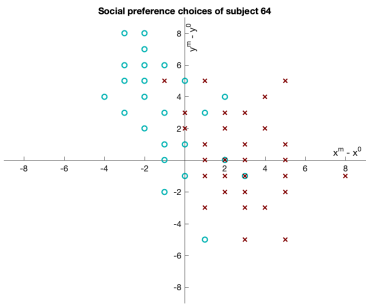
(c) Picks alternatives that maximise charity's payoffs



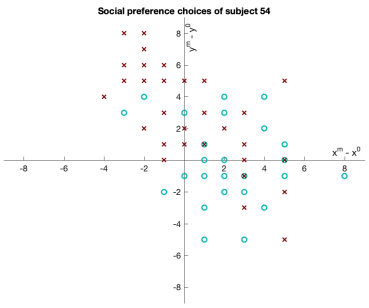
(d) Preferred alternatives from all regions



(e) Chooses mainly for DM's sake with more variation



(f) Non-systematic behaviour



Each panel of the figure shows a representative subject from each identified cluster in the social preference game. The x-axis displays $x^m - x^0$ and the y-axis displays $y^m - y^0$, with $m = 1$. A blue circle signifies that the subject chose the "reference point" and a red cross signifies that the subject chose the alternative option. The panels demonstrate high heterogeneity in preferences.

Cluster five contains only three subjects whose behaviour seems to be non-systematic (Figure 1.13(e)). Finally, cluster six contains only one subject who displays non-systematic behaviour (Figure 1.13(f)).

The behaviour displayed by the six groups is similar to that identified for the six groups of the lying game. Yet, the size of the respective groups differs between the two games. This suggests that subjects may swap their group membership and thus their behaviour across the games. Here, I assess this possibility at the group level. An analysis at the individual level is provided further below.

Above, we have seen that the behavioural clusters coincide in the identified behaviour across the lying and the social preference games. However, what we are interested in is not whether the same types of behaviour exist in general but whether the DMs stick to one behaviour across the two games or if they do not, whether the members of the clusters change their behaviour in the same way as the other members of their cluster. To assess this, we need to examine group membership in the clusters. If, for example, those DMs who maximised their own payoff in the lying game are not members of the cluster that maximises the DMs' payoffs in the social preference game, this suggests that DMs change their behaviour across games.

In order to assess whether this is the case, i.e. whether the same people are clustered together in the social preference game as in the lying game, I calculated the normalised mutual information (NMI) score. This score assesses to which percentage group membership coincides across the two games i.e. to which degree the two cluster analyses provide the same, thus mutual, information. The NMI score takes value one if all subjects who have been clustered into one group are also clustered into one group in the second sample. It takes value zero if group membership does not coincide at all. This measure takes into consideration label differences between analyses. For example, if a group of people is clustered into cluster "1" in the lie analysis and the same people are clustered into cluster "2" in the social preference analysis, this measure is able to identify that the people in both clusters are the same, despite being labelled differently. In such a

case, the score would be equal to one which would indicate perfect "stability".

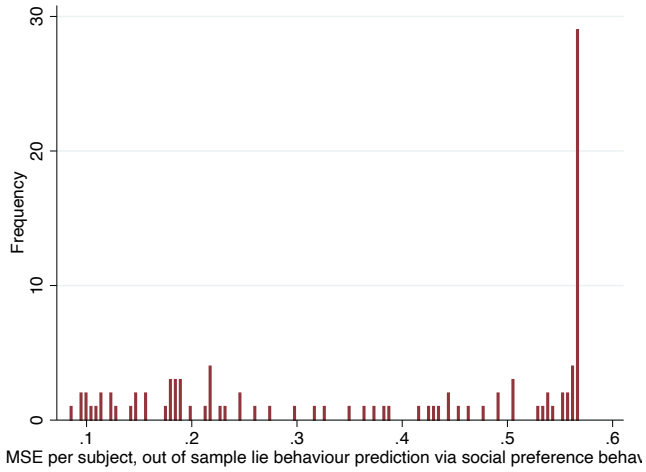
There exist three possibilities. First, DMs could behave in exactly the same way across both games. Second, DMs could change their behaviour but in such a way that most members of a cluster change their behaviour across games in the same way. Third, DMs might change their behaviour but members of a cluster in the the lying game may change their behaviour differently to each other. The first and second case would lead to a very high NMI score as people who have been clustered into one group in the lying game would also be clustered into one group, even if this describes different behaviour, in the social preference game. If, however, the NMI score is low, we would know that the third case is correct.

Running the analysis, I obtain an NMI score equal to 0.238. To compare, if group membership in the clusters from the lying game is compared to fully random clusters, the NMI score is equal to 0.11. If there were perfect correlation between two analyses instead, the score would be equal to 1. A score of 0.238 thus indicates that stability between the two cluster analyses is low and that we are consequently in the third case. This signifies that people who behave similarly in the lie experiment often do not behave similarly to each other in the social preference block and vice versa. Importantly, this implies that social preferences alone cannot explain lie preferences. Otherwise, being in one cluster in the social preference game should have had a higher correlation with behaviour in the lying game.

I further support the finding that lie preferences and social preferences do not seem to be systematically related by showing that one cannot use the behaviour from the social preference block behaviour to predict behaviour in the lie block well. To this end, I run an out-of-sample prediction analysis by individual to account for heterogeneity. I fit the extended benchmark model to the subject's social preference block behaviour and then use this fitted model to predict the subject's behaviour in the lying game. I repeat this for each subject in the sample. The mean squared forecast error is equal to 0.3834 which

is much higher than that in previous analyses. Its size implies that using a subject's social preference block behaviour to forecast their lie block behaviour is only a bit better than using a coin flip to predict their behaviour. Figure 1.14 shows a histogram of the mean squared forecast errors. Comparing this figure to the results in Figure 1.11 shows that the forecasting errors here are very large and that social preference game behaviour is thus not suitable to forecasting lying behaviour.

Figure 1.14: Performance of predicting lying behaviour based on social preference behaviour



Histogram of each subject's mean squared error (MSE) in out-of-sample prediction exercise. Behaviour in the social preference game is used to predict behaviour in the lying game.

In summary, while similar types can be classified in the social preference game as in the lying game, they are exhibited by different subjects across the two games. Importantly, subjects who share patterns of behaviour in one of the games mostly do not share patterns

of behaviour in the other game. Social preferences thus do not predict lying preferences.

1.5.2 Individual specific analysis

I now examine how subjects differ between the two blocks and, should there be no patterns across the whole sample, whether there are sub-group specific patterns. To examine these questions, I compare cluster membership in the social preference with that in the lie block subject by subject.

Never-liars, who form the largest lie block cluster, differ a lot from each other in the social preference block. Some of them behave to maximise their own payoff, others that of the charity and many to maximise a combination of the two payoffs. This is of great interest as these subjects acted to fulfil the moral code of "you should not lie". It would have been conceivable that they would act according to another moral code, such as "you should help others irrespective of the effect on yourself", in the social preference game. Instead, never-liars do not seem to share the same preferences in the social preference game.

When we analyse the other lie clusters, other interesting patterns emerge. For the lie cluster that maximises DM's payoffs, roughly two thirds also maximise the DM's payoff in the social preference block. However, the remaining one third behaves more egoistically in the lie block than in the social preference block where they maximise the combination of the DM's and charity's payoffs. For the lie cluster that maximises the combination of the DM's and the charity's payoffs, all subjects behave the same way in the social preference block. This implies that the extended benchmark model with a weakly positive cost of lying is sufficient to describe behaviour for this group of subjects. It also explains why the NMI score was low but larger than if the clusters had been randomly assigned: for this group, cluster membership coincides across the two games. For the lie block cluster that maximises the charity's payoff, a third of subjects also maximise the charity's payoff in the social preference block. The other two thirds

act more altruistically in the lie block than in the social preference block where they maximise the combination of the DM's and charity's payoffs. For the two lie block clusters that contain non-systematic behaviour, subjects' social preference block behaviour varies; some maximise the DM's payoff, some the charity's, some the combination of the payoffs and some behave non-systematically.

This reveals three patterns: A third of the subjects is consistent across the two blocks (32% of the whole sample); never-liars (43% of the whole sample) do not share the same social preferences among themselves, and there exist subjects who behave more altruistically and some who behave more egoistically in the lie block than in the social preference block (roughly 13% of the whole sample).[8] A narrative explanation for the latter two could be statements along the line: "If I lie, it should be to benefit someone else" and correspondingly, "If I lie, it should be worth my while and benefit me".

The results show that never-liars do not have common social preferences, and they therefore account for at least some of the variation in the group allocation between the lie and the social preference games. To check whether the low NMI score is driven by these never-liars, I reconducted the k-means analyses without the never-liars. For the analyses, I used $k = 5$ as one of the six groups identified in the initial k-means analysis of the lie game behaviour was formed of never-liars. Conducting the identical k-means analyses from Sections 1.4 and 1.5, including the same seed and PCA pre-step, but having dropped the never-liars beforehand and setting $k = 5$, yielded an NMI score of 0.2872. Recall that this score can be interpreted as the percentage of subjects who shared group membership across the lying and the social preference game, with the score equal to 1 if the groups perfectly coincide across games. This higher score, compared to the score of 0.238 for the analysis with never-liars included, confirms that part of the variation in the groups was driven by the never-liars. However, the score also shows that only around 30% of the remaining subjects

[8]The remaining 12% stem from subjects who are classified as non-systematic in the lie block and where systematic deviations are therefore difficult to assess.

share group membership in both games. This confirms that subjects' behaviour varies significantly between the lie and the social preference games.

In summary, both group and individual level analyses confirm that while social preferences enter lying preferences, they cannot fully capture them. The results thus demonstrate a need for theoretical frameworks that analyse lying preferences as something different from social preferences, while taking into account that social preferences matter to some degree, such as the framework introduced in this paper.

1.6 Parametrised and Calibrated Utility Function

Given that the previous sections have shown that behaviour across lie types is systematic, we can now exploit this for model building and thus for prediction. In this section, I present an example of a parametric utility function that fits the theoretical framework as well as the results from the experiment. This model can be estimated and can be used for model building in a variety of settings. While this is by no means the only possible utility function, it fits the data very well while being relatively simple, given the complex interaction between lie types and decision-maker types. The experimental results sections above have already shown that most subjects behave differently when they are in a distributive scenario without communication as compared to a lie setting with distributional effects. Therefore, the function presented here describes behaviour in the presence of lies, only.

1.6.1 Utility Function

I propose a utility function of the following shape:

$$u(s_d^m; s_d^0, \theta) = \max\{\alpha_\theta(x_d^m - x_d^0) + \beta_\theta(y_d^m - y_d^0), \delta_\theta(x_d^m - x_d^0) + \gamma_\theta(y_d^m - y_d^0)\} - \mathbb{1}_{s_d^m \neq s_d^0} c_\theta \tag{1.3}$$

Notice that $u(s_d^0; s_d^0, \theta) = 0$ and that therefore if $u(s_d^m; s_d^0, \theta) > 0$, then $r_d \neq s_d^0$ and else $r_d = s_d^0$. The following constraints ensure that the properties imposed by the framework hold:

Constraints:
$c_\theta \geq 0$
If Property 3(1) holds: $1 = \alpha_\theta = \delta_\theta, 0 = \beta_\theta = \gamma_\theta$
If Property 3(2) holds: $0 = \alpha_\theta = \delta_\theta, 1 = \beta_\theta = \gamma_\theta$
If Property 3(3) holds: $1 > \alpha_\theta = \delta_\theta > 0, \beta_\theta = \gamma_\theta = 1 - \alpha_\theta$
If Properties 3(1) & 3(2) hold: $\beta_\theta = 0, \delta_\theta = 0$
If Properties 3(1) & 3(3) hold: $\frac{\delta_\theta}{\gamma_\theta} > 0, \beta_\theta = 0$
If Properties 3(2) & 3(3) hold: $\frac{\alpha_\theta}{\beta_\theta} > 0, \delta_\theta = 0$

Specifically, the constraint on c_θ ensures that Property 2 holds for the utility function. The other constraints ensure that Property 3, the regions of inference property, holds. The properties thus influence strongly which behaviour is allowed within the framework: The properties provide structure on the utility function in the form of constraints.

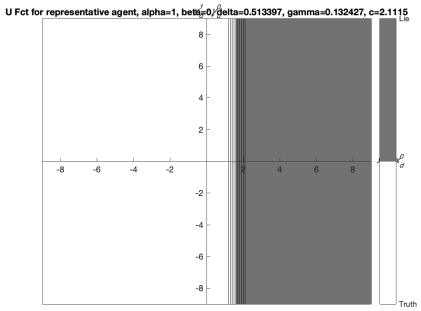
The utility function consists of a function of the relative payoffs, $v(\cdot)$, and a constant cost parameter. The $max\{\cdot, \cdot\}$ operator permits that behaviour varies across Δx^m and Δy^m and thus across lie types so that $v(\cdot)$ also depends on the lie type. Parameters $\alpha_\theta, \beta_\theta, \delta_\theta, \gamma_\theta$ determine to which degree the decision-maker cares about their own payoff compared to the partner's payoff. If $\alpha_\theta = \delta_\theta$ and $\beta_\theta = \gamma_\theta$, the model collapses to the extended benchmark model that was introduced earlier. Due to the $max\{\cdot, \cdot\}$ operator, the degree to which the decision-maker cares about the payoffs can vary across the payoff space. The approach is similar to piece-wise linearity.

1.6.2 Calibration

I first adopt a representative agent approach and in a second step demonstrate that we should consider heterogeneity of decision-makers to model behaviour.

First, I calibrate the model to describe the representative agent. To this end, I perform an MLE estimation with a probit likelihood for each of the subjects and then calculate the median estimates. This yields behaviour as shown in Figure 1.15. The figure shows estimated lie and truth-telling regions (lie regions in grey, truth-telling regions in white) for the representative agent with the gain from lying to the DM shown on the x-axis and that of the charity on the y-axis.

Figure 1.15: Estimated lie regions for the representative agent of the lying game



The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Dark grey areas indicate that the agent is expected to lie and white areas indicate payoff combinations for which the agent is expected to tell the truth.

Figure 1.15 suggests that DMs only care about their own payoffs and start to lie as soon as their payoff gain from lying, $x_d^m - x_d^0$, is

larger than a positive constant, the constant cost of lying. The figure shows a very distinctive pattern of behaviour. However, the previous sections have shown that lying behaviour is highly heterogeneous. I therefore repeat the exercise for each of the clusters from Section 1.4. I thus obtain six representative agents whose behaviour can be compared.

This methodology has the advantage that the MLE estimates themselves are not conditioned on the clusters as the model is calibrated to each DM individually. Therefore, no prior on group membership enters the calibration process, reducing the chance of bias in the estimates. The median representative agent of each cluster is then obtained by taking the median of the parameters per cluster.

Table 1.4 shows the parameter estimates for the representative agents across clusters as well as for the whole sample. It is important to note that because the median value of each parameter is shown, the constraints might not hold for all rows in the table.

Figure 1.16 shows the estimated lie (grey) and truth (white) regions for the estimated representative individuals for the clusters defined in Section 1.4 above.[9]

Never-liars' estimates are characterised by taking a large value for c_θ and other parameter estimates are either equal to zero or small relative to the constant cost. The combination of the parameters ensures that these subjects do not lie in the payoff space of the experiment.

[9] As an alternative approach to the median representative agents one can calculate the median behaviour of each cluster, thereby obtaining behaviourally representative agents. The model is then calibrated to each of these six behaviourally representative agents. While the resulting estimates are more representative of the clusters, the approach suffers from a sequential testing problem where, if the cluster allocation itself were flawed, the results of the calibration would be as well. For completeness, I provide the results of this alternative methodology in Section 1.8.6 in the Appendix. The results are very similar to each other.

Table 1.4: Parameter values by representative agent for the full sample and by group

Type, θ	α_θ	β_θ	δ_θ	γ_θ	c_θ
Full sample	1	0	0.51	0.13	2.11
	(0.48)	(0.48)	(0.61)	(0.34)	(1.66)
Never-liar	1	0	1	0	22.20
	(0.04)	(0)	(1)	(0)	(18.62)
Max. DM's payoff	0.86	0.14	0.76	0.14	0.64
	(0.21)	(0.21)	(0.20)	(0.20)	(1.30)
Max. both payoffs	0.52	0.48	0.52	0.48	0
	(0.07)	(0.07)	(0.08)	(0.07)	(0.09)
Max. charity's payoff	0.40	0.61	0.05	0.65	0.81
	(0.41)	(0.72)	(0.29)	(0.50)	(0.97)
Balancing behaviour	0.88	0	0	0.26	1.63
	(0.55)	(0)	(0)	(0.12)	(0.86)
Non-strategic behaviour	1.11	0	0	0.15	1.72
	(0.59)	(0)	(0)	(0.32)	(1.00)

The median parameter values calculated by cluster are given for the whole sample and by cluster. The interquartile range for each parameter is shown in brackets under the median.

For those subjects that lie mainly when the DM gains, subjects are characterised by relatively small parameter estimates that, in combination, put the most weight on the DM's monetary payoff.

For those DMs who lie for their own as well as for the charity's gain parameter estimates are also small. However, the constant cost is close to zero and the weights on $x_d^m - x_d^0$ and $y_d^m - y_d^0$ are nearly equal for both functions of the max.

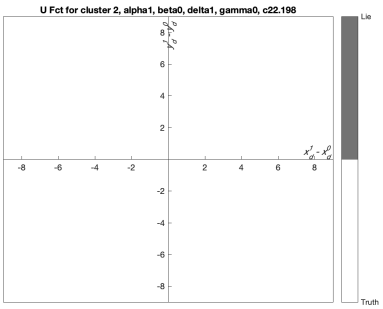
For those who lie mainly to benefit the charity, parameter estimates are similar in size to those for the previous cluster. However, as expected, the weights on the charity's payoff are larger than those on the DM's. Perhaps surprisingly, the weight on the DM's payoff is very small for the second function in the max operator. This implies that there is a kink in the lie region, which is shown in Figure 1.16(d).

For the two groups where behaviour looks non-systematic, the suspected balancing and the non-systematic group, the estimation procedure indicates that Properties 3(1) and (2) hold simultaneously and that there is thus a kink and that behaviour is described by Property 3(1), respectively.

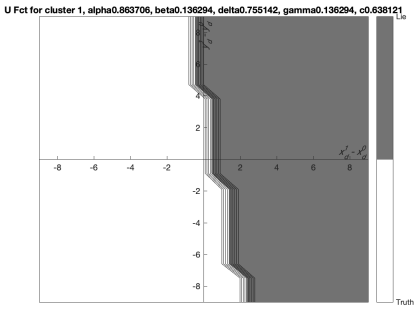
These lie regions provide a simple tool for the analysis and prediction of lying behaviour in binary choice settings. Based on the estimation results, such a figure can be created for every subject in the dataset. Due to the single-crossing property, Property 3, imposed by the framework it is possible to visualise how a subject is expected to behave for alternatives that were not included in the experiment. The lie regions thus provide the researcher with a visual tool to help anticipate behaviour without requiring additional data.

Figure 1.16: Estimated lie regions for representative agents of groups identified in Section 1.4

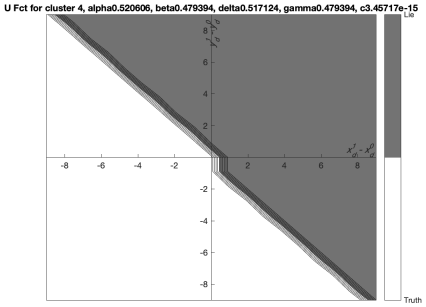
(a) Never-liar



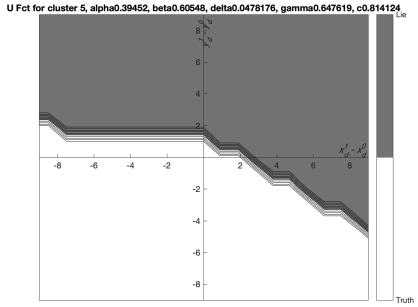
(b) Lies mainly when DM gains



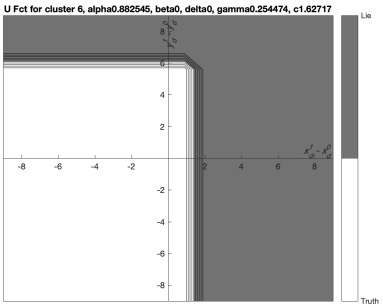
(c) Lies for both DM's and charity's gain



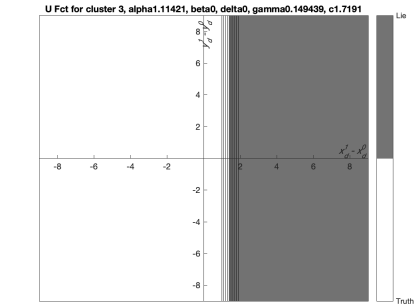
(d) Lies when the charity gains



(e) Balancing?



(f) Non-strategic behaviour



The x-axis displays $x_d^1 - x_d^0$ and the y-axis displays $y_d^1 - y_d^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Dark grey areas indicate that the agent is expected to lie and white areas indicate payoff combinations for which the agent is expected to tell the truth.

1.7 Conclusion

This paper proposes a unifying framework in which lying preferences can be analysed. The framework defines the setting and the space of lie types and provides properties of behaviour. It also informs the design of an experiment aimed at eliciting lying preferences in this setting. The results show that the identified lie types matter and that the properties impose plausible constraints on behaviour. Importantly, accounting for both improves predictive power relative to existing benchmark models. The experimental component introduces an experimental design that models lie types in a straightforward manner, permits the researcher to observe lying choices at the individual level and allows the clean separation of lying and social preferences. The design can easily be paired with modern machine learning methods which are useful for the analysis of decision-maker types. Employing a combination of principal component analysis and a k-means algorithm, I find that there are six major behavioural types of decision-makers. Knowing to which group a decision-maker belongs vastly improves the performance of out-of-sample predictions. In contrast, knowing how the subject behaved in a social preference game analogous to the lying game does not help to predict lying decisions. Finally, I propose a parametric model and calibrate it to describe behaviour. The need for a unifying framework of lying is underlined by the large improvements in predictive accuracy when accounting for the heterogeneity of lies and decision-makers that are the key elements of the framework.

To conclude, in this paper, I have examined the fundamentals of lying preferences in the presence of heterogeneity of lie types and decision-maker types. The resulting insights can be used for the purpose of model building and prediction analyses. Having identified and modelled the fundamental preferences, the next step is to establish their relationship with other aspects that enter the decision to lie such as time pressure concerns or reputation and probability of detection which are important in repeated interactions.

Acknowledgements:
I would like to thank Larbi Alaoui and Jose Apesteguia for their invaluable comments, support and advice. I would also like to thank Antonio Penta for his support and helpful comments. Further, I thank Johannes Abeler, Pierpaolo Battigalli, Antonio Cabrales, Colin Camerer, Francesco Cerigioni, Andrew Ellis, Lukas Hoesch, Navin Kartik, Gaël Le Mens, Rosemarie Nagel, Yusufcan Masatlioglu, Pedro Rey Biel and Balazs Szentes as well as seminar participants at the 2020 Econometric Society European Winter Meetings, the 2020 Barcelona GSE PhD Jamboree, at Universitat Pompeu Fabra and the London School of Economics for their useful feedback and comments.

1.8 Appendix

1.8.1 Instructions of the Experiment

1.8.1.1 Introduction

Thank you for deciding to participate. The experiment is split into two large blocks of questions. Each block contains 60 short rounds. Your choices in the experiment will affect your bonus payments and that of a partner. Specifically, this partner is a charity that you can select from the list below. At the end of the experiment, one question from each block will be selected for payment purposes by chance. You will receive the payment as a bonus payment via Prolific and the charity will receive the payment through a donation. Payoffs will be displayed in terms of tokens. These tokens will be converted into GBP at the end of the experiment. 5 tokens correspond to GBP1. Before you start each block, you will receive detailed instructions for that block.

Which charity would you like to be partnered with? Please select one.

Please select a charity before you continue.

- Cancer Research UK
- Children in Need
- Comic Relief
- National Trust
- WWF

1.8.1.2 Instructions - Lying game

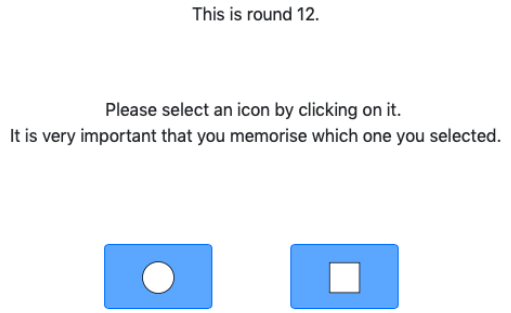
Please read the instructions carefully.

In this block, you will play 60 rounds of the following format: Each round consists of 2 screens, displayed directly after one another.

First screen:
You have to select one icon by clicking on the button on which the icon is displayed.
There exist 6 icons, each a geometrical shape, in total: square, circle, triangle, diamond, pentagon and hexagon.
It is very important that you memorise which icon you have selected.

Below, you can see an example of the first screen:

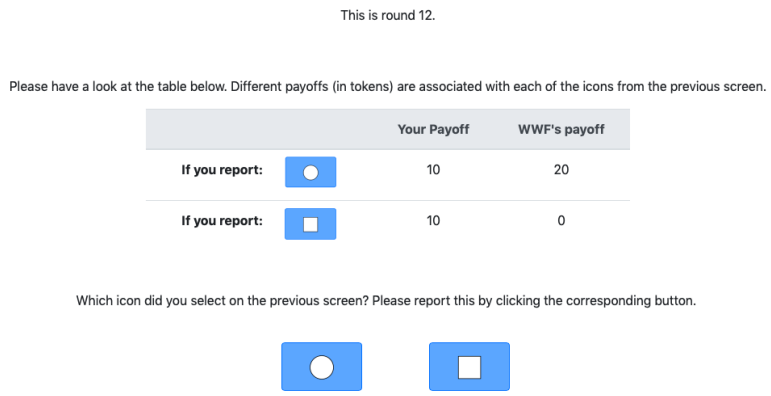
Figure 1.17: Example screen of the 1st screen of one of the rounds of the lying game.



Second screen:
After having selected the icon on the previous screen, you will see a screen similar to the one below:

To each icon from the first screen correspond 2 token values.

Figure 1.18: Example screen of the 2nd screen of one of the rounds of the lying game.



In the table, you can see the token values of each icon by looking at the row that shows the respective icon.
The first column shows your monetary value and the second column the charity's monetary value.
Importantly, the icons that are displayed on the buttons are drawn randomly each round which means that the token values associated with a particular icon vary each round.

Decision:
At the bottom of the second screen, you have to report which icon you selected in the first screen.
Your payoff from this round will be determined by the icon you report on the second screen and not the one you selected on the first screen. We understand that each round, you may have reasons for or against reporting truthfully.

Payments:
In this block, one round will be selected by chance. You and the charity that you selected will receive the payments according to the

button that you reported in this round.

Please answer the following questions to confirm that you have understood the instructions.

You have to answer all questions correctly before you can continue with the experiment.

1) Imagine that you selected the square on the first screen (in the image above). Then, on the second screen (in the second image above), you clicked on the circle. Which of the following statements is true?

- I told the truth when I reported the circle and the charity will receive a larger payment than if I had clicked on the square.
- I told the truth when I reported the circle and the charity will receive a smaller payment than if I had clicked on the square.
- I told a lie when I reported the circle and the charity will receive a larger payment than if I had clicked on the square.
- I told a lie when I reported the circle and the charity will receive a smaller payment than if I had clicked on the square.

2) With whom will you be matched during the experiment?

- A charity of my own choosing
- A computer generated player

3) What will your payment depend on?

- The icon that you selected on the first screen
- The icon that you reported on the second screen

1.8.1.3 Instructions - Social preference game

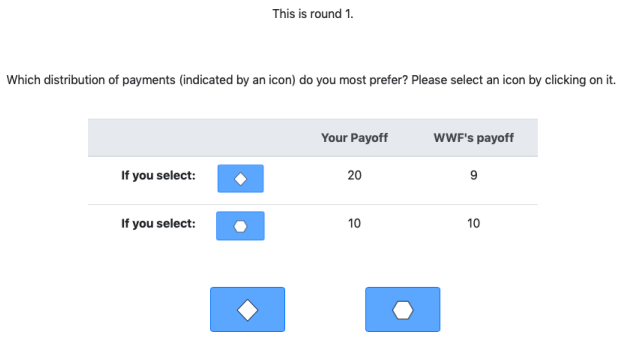
Please read the instructions carefully.

In this block, you will play 60 rounds where each round consists of exactly one question screen.

Question screen:
You will see a table where each row corresponds to a choice object and each column to monetary payoffs (in tokens).

The first column shows your monetary gain from selecting that icon and the second column the charity's monetary gain. Please take a look at the example screen below:

Figure 1.19: Example screen of one of the rounds of the social preference game.



Decision:
At the bottom of the screen, you have to select which of the options i.e. icons you prefer.
Your payoff and that of the charity from this round will be determined by the icon that you click on.
There exist 6 icons, each a geometrical shape, in total: square, circle,

triangle, diamond, pentagon and hexagon.
The symbols are allocated randomly to a payoff each question. For example, that means that the payoff pair "10 for you and 5 for the charity" could have any of the icons next to it.

Payments:
In this block, one round will be selected by chance. You and the charity that you selected will receive the payments according to the icon that you selected in this round.

Please answer the following questions to confirm that you have understood the instructions.

You have to answer all questions correctly before you can continue with the experiment.

1) With whom will you be matched during the experiment?

- A charity of my own choosing
- A computer generated player

2) In the example above, which choice will maximise your payoff?

- Clicking on the button displaying the diamond
- Clicking on the button displaying the hexagon

3) In the example above, which choice will maximise the charity's payoff?

- Clicking on the button displaying the diamond
- Clicking on the button displaying the hexagon

1.8.2 Specification of rounds of the experiment

Table 1.5: Rounds with non-binary choice

Types of lies	(x^0, y^0)	(x^1, y^1)	(x^2, y^2)	(x^3, y^3)	(x^4, y^4)	(x^5, y^5)
MHL	(5, 6)	(4, 4)				
HL	(5, 6)	(5, 5)				
SHL	(11, 10)	(10, 10)				
WAL	(10, 10)	(10, 13)				
WAL	(10, 8)	(10, 13)				
WAL	(10, 9)	(10, 10)				
WAL	(10, 11)	(10, 13)				
SSL	(10,10)	(11, 10)				
SSL	(10,10)	(15,10)				
SSL	(10, 9)	(12, 9)				
SSL	(7, 10)	(10, 10)				
MBL	(14,12)	(15, 15)				
MBL	(10, 8)	(14, 10)				
MBL	(10, 8)	(11, 9)				
MBL	(8, 10)	(11, 11)				
MBL	(7, 9)	(12, 10)				
MBL	(7, 10)	(9, 12)				
MBL	(8, 10)	(10, 13)				
MBL	(10, 9)	(12, 13)				
MBL	(9, 7)	(10, 12)				
MBL	(6, 10)	(10, 14)				
MBL	(7, 7)	(10, 10)				
EL	(10, 10)	(12, 8)				
EL	(10, 8)	(11, 7)				
EL	(10, 15)	(11, 10)				
EL	(10, 15)	(15, 10)				
EL	(9, 12)	(10, 9)				
EL	(9, 10)	(12, 9)				
EL	(9, 12)	(11, 11)				

EL	(10, 10)	(13, 8)				
EL	(7, 13)	(10, 10)				
EL	(7, 11)	(12, 10)				
EL	(6, 14)	(9, 9)				
EL	(11, 10)	(15, 7)				
EL	(2, 10)	(7, 8)				
EL	(5, 10)	(13, 9)				
AL	(10, 10)	(9, 11)				
AL	(10, 5)	(9, 8)				
AL	(10, 9)	(9, 15)				
AL	(10, 6)	(8, 8)				
AL	(15, 5)	(13, 13)				
AL	(10, 10)	(8, 14)				
AL	(14, 8)	(12, 13)				
AL	(10, 10)	(7, 13)				
AL	(12, 4)	(10, 10)				
AL	(14, 6)	(10, 10)				
AL	(15, 5)	(12, 11)				
AL	(14, 7)	(11, 12)				
AL	(15, 4)	(12, 12)				
AL	(16, 5)	(14, 12)				
MBL, EL	(10, 10)	(15, 15)	(18, 9)			
MBL, EL	(10, 10)	(11, 11)	(13, 9)			
AL, EL	(10, 10)	(9, 15)	(15, 9)			
EL, AL	(12, 10)	(15, 9)	(11, 15)			
SSL, WAL	(10, 10)	(15, 10)	(10, 15)			
SSL, WAL	(5, 10)	(10, 10)	(5, 11)			
WAL, AL	(10, 10)	(10, 12)	(9, 15)			
SSL, EL	(10,10)	(12, 10)	(15, 9)			
MBL, SSL, WAL, AL, EL	(10, 10)	(11, 11)	(12, 10)	(10, 12)	(9, 15)	(15, 9)
MBL, SSL, WAL, AL, EL	(8, 10)	(11, 11)	(8, 15)	(6, 18)	(13, 10)	(16, 8)

The table displays the payoff bundles for all experimental rounds with more than

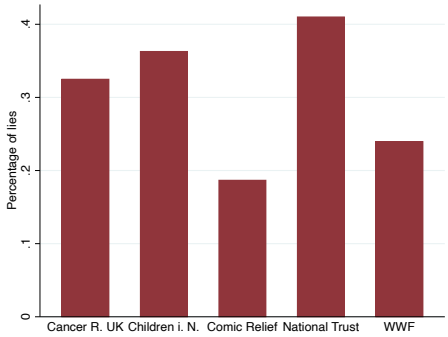
two states. (x^0, y^0) refers to the payoffs of the true state; all other payoff bundles are linked to untrue states that are available for reporting in addition the true state. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *egoistic lie* (EL), *self-serving lie* (SSL), *self-harming lie* (SHL), *harmful lie* (HL) and *mutually harmful lie* (MHL).

1.8.3 Details on comparison of aggregate results to the literature

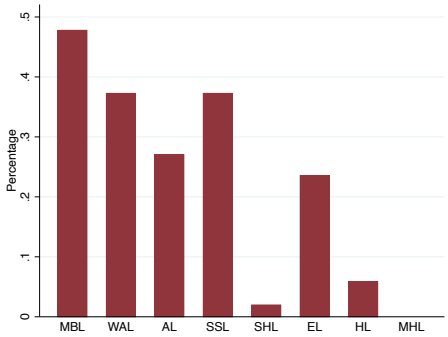
1.8.3.1 Lying behaviour across charities

Figure 1.20: Average percentage of lies by lie type for each charity

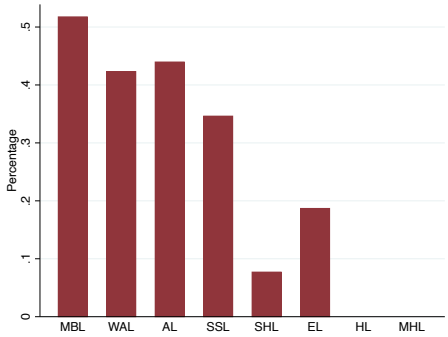
(a) Average number of lies by charity



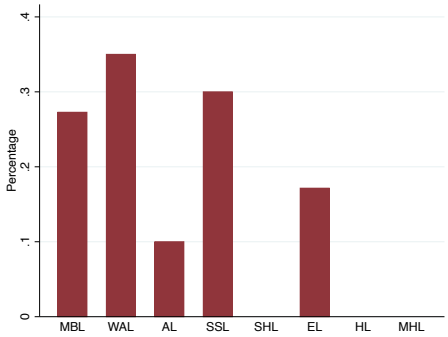
(b) Cancer Research UK



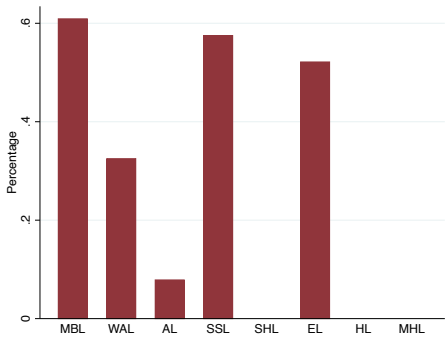
(c) Children in Need



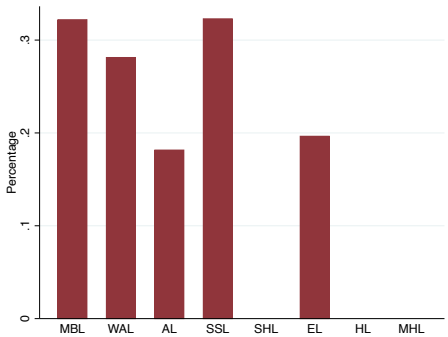
(d) Comic Relief



(e) National Trust



(f) World Wildlife Fund



Panel (a) shows the average percentage of lies told by subjects who selected the charity. Each bar represents one charity. Panels (b) - (f) show the percentage of lies by lie type for for each of the charities. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *self-serving lie* (SSL), *self-harming lie* (SHL), *egoistic lie* (EL), *harmful lie* (HL) and *mutually harmful lie* (MHL).

1.8.3.2 Covariates of lying

After completion of the two main stages of the experiment, subjects were asked to answer a series of questions and to perform cognitive ability as well as personality tests. The following paragraphs examine whether there exists a relationship between certain personality or cognitive traits as well as demographics and lying behaviour. Results are displayed in Table 1.6.

A commonly discussed feature in the literature is gender. I find a significant difference in the average percentage of lies between men and women (t-test, p-value¡0.01) overall as well as for most of the lie types. Specifically, I find that men lied significantly more (36.33% compared to 26.23% across all questions). The same relationship exists for *mutually beneficial lies (MBLs)*, *weakly altruistic lies (WALs)*, *self-serving lies (SSLs)* and *egoistic lies (ELs)*. Not surprisingly, there is no difference for *self-harming lies (SHLs)*, *harmful lies (HLs)* and *mutually harmful lies (MHLs)* as nearly no subject lied for these. Interestingly, there exists no significant difference in the behaviour for *altruistic lies (ALs)*. The finding that men lie more in MBLs is in line with Erat and Gneezy (2012) but goes against the finding of no differences in Biziou-van Pol et al. (2015) and Cappelen et al. (2013). The finding that more men lie for ELs is in line with Dreber and Johannesson (2008) while the finding of no differences for ALs goes against Erat and Gneezy (2012)'s finding that more women tell ALs. Yet, it is noteworthy that ALs are the only type of lies for which women lied more than men (except for HLs and SHLs which are likely to be due to errors) which, while statistically insignificant, does provide some evidence in the direction of Erat and Gneezy (2012), especially when comparing this to the large differences between men and women's percentage to lie for the other lie types. Importantly, previous papers often asked a very low number (sometimes only one) of lie questions so that it is perhaps not surprising that those results are volatile when using different question specifications.

A second feature of interest is cognitive ability. Subjects answered a CRT test (see Thomson and Oppenheimer (2016) and Frederick (2005)) where they received one point for each correct answer out of a total of four questions. Testing lie behaviour for each level of the score as well as for measures that split the sample into two groups based on their score (two different splits were used) reveals that subjects with different scores answered questions significantly differently both on average and for each of the lie types. Interestingly, subjects with a higher CRT score lied significantly more for each lie type except for HLs where subjects with lower scores lied more (this is likely to be caused by choice errors of which subjects with low CRT scores seem to perform more).

Subjects also answered a ten item Big Five personality test (Rammstedt and John (2007)). Splitting subjects into two groups (high versus low performing) for each of the big five character traits, I analyse if there exist systematic differences. For example, subjects who scored more than five out of ten points on the conscientiousness measure lied significantly less than people who scored five or fewer points. This also holds when looking at MBLs, WALs, SSLs and ELs, specifically. However, there is no difference in behaviour for ALs. Conscientiousness could be linked to being more rule-abiding. This could then explain the lower number of lies even when these lies help the charity.

Table 1.6: Significant differences in group behaviour by covariates of interest

Covariates	Average	MBLs	WALs	ALs	SSLs	SHLs	ELs	HLs
Gender	***	***	***		***		***	
Male	36.33%	54%	42.5%	24.14%	46.5%	0%	30.43%	2%
Female	26.23%	36.36%	28.3%	24.66%	28.77%	3.78%	18.87%	3.78%
CRT	***	***	***	***	**		***	
2 or fewer points	25.11%	35.40%	27.13%	18.54%	30.32%	2.13%	21.43%	6.38%
3 or 4 points	36.19%	52.92%	41.96%	29.34%	43.30%	1.79%	27.04%	0%
Conscientiousness	***	***	***		***		***	
5 or fewer points	40.40%	57.14%	42.86%	25.51%	54.76%	4.76%	38.10%	0%
6 to 10 points	28.76%	41.80%	33.23%	24.13%	32.93%	1.22%	20.99%	3.66%

Group behaviour is defined as the percentage of lies for each lie type of a group. Stars indicate significance levels with p¡0.1 *, p¡0.05**, p¡0.01***. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *self-serving lie* (SSL), *self-harming lie* (SHL), *egoistic lie* (EL) and *harmful lie* (HL). *Mutually harmful lies* are not displayed as the percentage of lies was 0% for all comparisons.

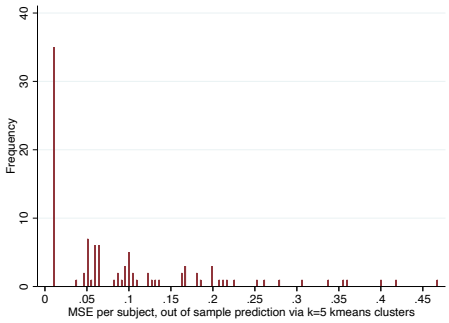
1.8.4 Supporting results for Section 1.4

1.8.4.1 Robustness to misspecification of number of clusters k

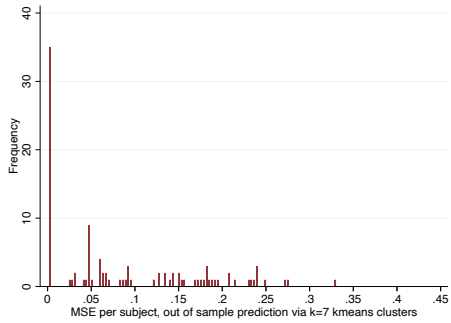
The figure shows the results from rerunning the heterogeneity out-of-sample forecasting exercise with $k = 5$ and $k = 7$. The comparison to Figure 1.11(b) shows that the gains from accounting for heterogeneity in lying preferences are robust to slight misspecification of k.

Figure 1.21: Histogram of each subject's MSE in out-of-sample prediction exercise with heterogeneity in preferences taken into account for different k.

(a) $k = 5$ clusters



(b) $k = 7$ clusters



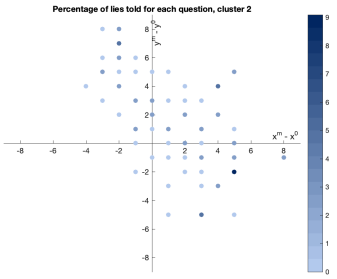
1.8.4.2 Never-liars and potential errors of behaviour

The k-means algorithm clustered 44 subjects into the never-liar group. Of these, 28 never misreported. The remaining subjects misreported one to 7 times. While I cannot rule out alternative explanations, these misreports appear to be honest mistakes rather than lies. First, shifts appear to be random rather than systematic even examining alternative explanations such as inequality concerns. Second, the k-means algorithm identified them as never-liars.

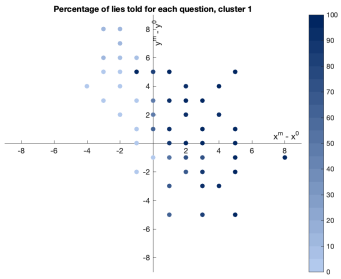
1.8.4.3 Behaviour of clusters

Figure 1.22: Percentage of lies told for each question in the lying game by cluster

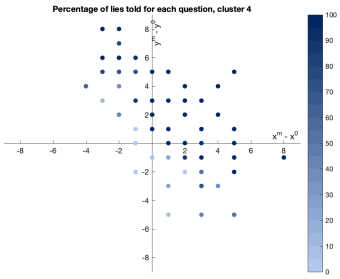
(a) Never-liar



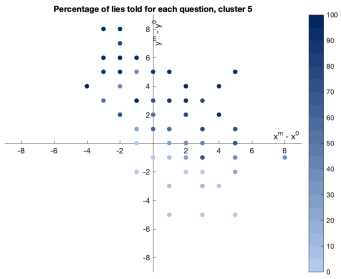
(b) Lies mainly when DM gains



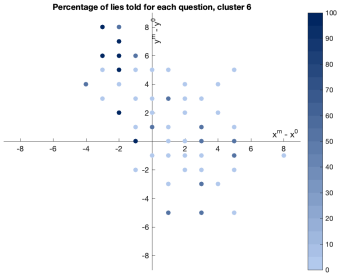
(c) Lies for both DM's and charity's gain



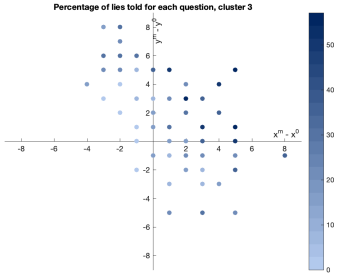
(d) Lies when the charity gains



(e) Balancing?



(f) Non-systematic behaviour

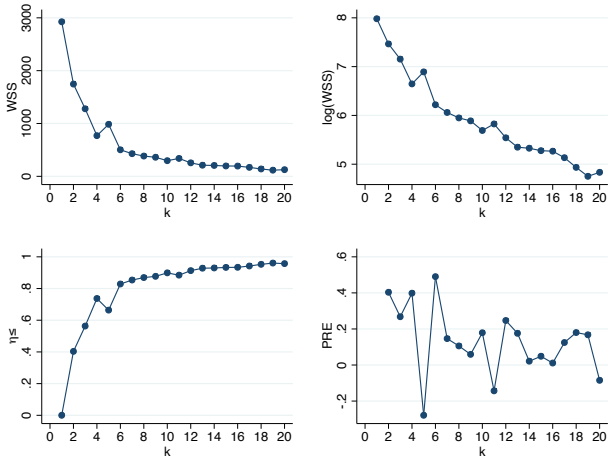


The x-axis displays $x^m - x^0$ and the y-axis displays $y^m - y^0$, with $m = 1$. Each dot represents a round with a binary choice of the lying game expressed via the payoff combinations. Lighter blues indicate a higher percentage of truths told and darker blues a higher percentage of lies told. The color scale to the right links the percentage to the colour.

1.8.5 Supporting results for Section 1.5

For robustness of the k-means analysis, I conducted a pre-analysis to identify the ideal number of k. The results here are not as clear as in the lie k-means analysis as explanatory power increases when having more than six groups. However, even with six groups, there are two groups that contain between one and three subjects which indicates that the clustering is already quite detailed. For that reason, it seems that six groups are satisfactory. Figure 1.23 shows summary statistics for the performance of 1 to 20 clusters.

Figure 1.23: Performance indicators by cluster number in social preference game k-means analysis



Plot of weak sum of squares (WSS), log WSS , η^2 and proportional reduction of error (PRE) for number of clusters $k = \{1, ..., 20\}$.

To rule out that misspecification of the number of clusters is driving the low NMI score, I reran the analysis with $k = 4$ clusters.[10]

[10]The fact that two groups have only between one and three members makes

Specifically, the k-means analyses from Sections 1.4 and 1.5 were repeated with k=4 and the NMI score was calculated for the new clusters. The resulting NMI score was equal to 0.1998, which is very similar to the NMI score obtained from comparing the clusters of the k-means analyses with $k = 6$. This suggests that the finding that cluster membership does not overlap between the lying and the social preference games is robust to the number of clusters specified and thus increases confidence in this finding.

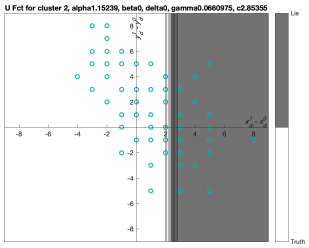
1.8.6 Supporting results for Section 1.6

Alternative calibration exercise.

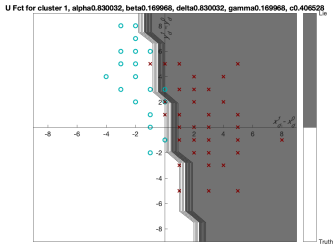
$k = 4$ the most likely contender.

Figure 1.24: Estimated lie regions for representative agents based on the mean behaviour of each cluster identified in Section 1.4

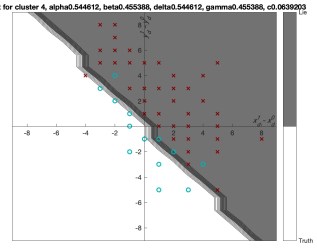
(a) Never-liar



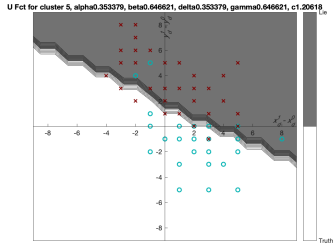
(b) Lies mainly when DM gains



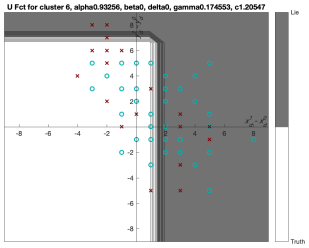
(c) Lies for both DM's and charity's gain



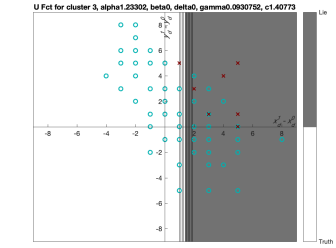
(d) Lies when the charity gains



(e) Balancing?



(f) Non-systematic behaviour



The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Dark grey areas indicate that the agent is expected to lie and white areas indicate payoff combinations for which the agent is expected to tell the truth.

1.8.7 Glossary for machine learning methodology

This section provides definitions of the terminology used in Sections 1.4 and 1.5 to describe the machine learning (ML) methodology used in the paper.

- *Label:* The dependent variable.
- *Unsupervised learning:* A ML algorithm that transforms data (for example predicts or groups data) where the true labels, i.e. the value of the dependent variable, are unknown to the researcher. One can imagine the difference between supervised and unsupervised learning to be parallel to within versus out-of-sample predictions. In sample, the value of the dependent variable is known so that the researcher can compare the performance of the prediction against the true values; this is similar to supervised learning. Out-of-sample, the value of the dependent variable is forecasted but is unknown to the researcher; this is akin to unsupervised learning.
- *Cluster:* A group of observations that are most similar to each other.
- *Centroid:* A centroid is the center of a cluster. This location does not have to be the real center as it is often initially allocated randomly.
- *k-means clustering:* An unsupervised ML algorithm where data is sorted into k clusters. The algorithm is initially allocated random centroids for each cluster. Based on the distance of a set of pre-specified independent variables (also called features in the ML literature) to the centroid, observations are allocated to each cluster. The centroids are then updated in that they are moved to the mid position of the features of the observations that had been allocated to the cluster in the previous step. The observations are then reallocated based on which of the updated

centroids is the closest to them. This procedure is repeated until updating the centroids does not lead to a change in allocations any more.

- *Principal component/factor:* A weighted, often linear, combination of correlated independent variables.
- *Principal components analysis (PCA):* The (unsupervised) process of identifying the principal components/ factors in the data. As the components are ordered by how much of the variance in the data they can explain, they can be used to reduce the number of variables needed in an analysis.
- *Factor model analysis:* The process of identifying the principal components/ factors that describe the most variance in the data and restricting the model to these factors.

Chapter 2

ELICITING PREFERENCES FOR TRUTH-TELLING IN A SURVEY OF POLITICIANS

(joint with Aina Gallego)

This paper has been published in the Proceedings of the National Academy of Sciences in 2020, see the full citation below:

Janezic, K. A., & Gallego, A. (2020). Eliciting preferences for truth-telling in a survey of politicians. Proceedings of the National Academy of Sciences, 117(36), 22002-22008.

URL: https://www.pnas.org/content/117/36/22002

DOI: https://doi.org/10.1073/pnas.2008144117

Significance Statement: Voters who would like to accurately evaluate the performance of politicians in office often rely on incomplete information and are uncertain whether politicians' words can be trusted. Honesty is highly valued in politics because politicians who are averse to lying should in principle provide more trustworthy information. Despite the importance of honesty in politics, there is no scientific evidence on politicians' lying aversion. We measured preferences for truth-telling in a sample of 816 elected politicians and study observable characteristics associated with honesty. We find that in our sample, politicians who are averse to lying have lower reelection rates, suggesting that honesty may not pay off in politics.

2.1 Introduction

A common stereotype across countries and time is "all politicians are liars". Politicians often face incentives to lie rather than tell the truth, for instance when damaging information can be hidden or undeserved credit can be claimed, while voters need accurate information to hold them accountable. Because lies in politics are hard to detect, politicians' dishonesty makes it difficult for voters to evaluate their performance. The problem of lies in politics is old, but the rise of fake news and posttruth politics has recently revived concern (Grinberg et al. (2019), Bovet and Makse (2019), Vosoughi et al. (2018)). In principle, the prevalence of lies in politics, and the ensuing distrust, could be reduced if politicians in office were averse to lying. Indeed, honesty is often considered one of the most desirable traits in politicians because it provides an internal drive to adhere to ethical behavior even when such behavior is invisible to others (Fearon (1999), Besley (2005), Caselli and Morelli (2004)). Yet, voters trying to tell honest and dishonest politicians apart face a vexing problem. Since politicians who bluff, displace blame, or use strategic deception try to appear honest, identifying those who are dishonest is extremely challenging.

Despite the importance of honesty in politics, sound empirical evidence about the observable correlates of preferences for truth-telling among politicians is lacking. A rapidly growing literature in behavioral economics and social psychology studies preferences for truth-telling (also called lying aversion or intrinsic honesty in the literature) in the general population by devising behavioral games that incentivize lying (Fischbacher and Föllmi-Heusi (2013), Gneezy et al. (2013), Gächter and Schulz (2016)). Some empirical studies have used such behavioral instruments to study honesty in populations that are both powerful and burdened by concerns about the integrity of their members such as the banking profession (Cohn et al. (2014)). Yet, to our knowledge, no studies to date have used behavioral instruments to measure honesty in samples of political elites.

Supporting the intuition that some people are more honest than others, research about preferences for truth-telling finds clear individual differences in the disposition to lie (Gächter and Schulz (2016), Abeler et al. (2019)). While situational elements affect lying behavior (Capraro et al. (2019), McLeod and Genereux (2008)), some people have a consistent preference for truth-telling even when lying is personally beneficial and not observable to others.

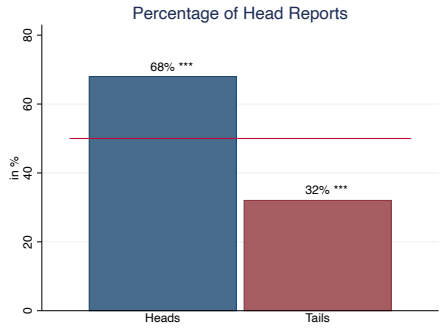
This paper studies truth-telling among politicians using a lying game with a nonmonetary incentive. We define a lie as misreporting private information and design a game in which politicians must flip a coin and have incentives to report heads. The literature defines several different types of such lies (Erat and Gneezy (2012)). This paper focuses on one specific type: nonobservable lies that only benefit the liar. Politicians face many situations in which they have private information, in which it is beneficial to them to be dishonest and in which their dishonesty has a low chance of being discovered. For example, politicians have been known to bury or misrepresent reports that do not align with their policy goals (see for example Maravall (1999)). This constitutes dishonest behavior as the population is only allowed to see information that is in line with policy goals, leading to misdirection and making it difficult for the electorate to evaluate

politicians' performance and thus undermining accountability. We focus our analysis on lies that are typical of such situations.

The lying game was embedded in a large survey of 816 Spanish mayors of municipalities with more than 2,000 inhabitants conducted between July 2018 and January 2019. In standard behavioral games, lying is incentivized by conditioning a monetary compensation on obtaining a specific outcome in a luck task such as rolling dice or flipping coins (Fischbacher and Föllmi-Heusi (2013), Abeler et al. (2014)). However, a pilot study revealed that monetary incentives are inappropriate in a study of professional politicians because they become alarmed or offended by the association of offering money with accusations of corruption (see Materials and Methods). Instead, we incentivized lying with a nonmonetary reward, a personalized report containing the results of the survey, which was highly valued by our sample. We recorded interest in receiving such a report at the start of the survey. At the end of the survey, we told mayors that they would only receive the report if they obtained heads in a private coin flip. As 88% of mayors were interested or very interested in receiving the report, they had an incentive to lie about the outcome of the coin flip. Because lying is incentivized effectively and reputational concerns are eliminated by the impossibility to tell if a particular politician lied, differences between subgroups in the propensity to report heads can be attributed to differences in preferences for truth-telling. Coin flipping and die rolling tasks have been shown to be valid measures of dishonesty as behavior in those tasks has been found to correlate with real-world measures of dishonesty such as avoiding paying for a ticket on public transport or not returning money when being overpaid (Gächter and Schulz (2016), Cohn and Maréchal (2018), Dai et al. (2017), Hanna and Wang (2017), Potters and Stoop (2016)).

Using this design with nonmonetary incentives, we first discover that a large and statistically significant proportion of mayors lied. In fact, they lied more often than other populations previously studied using similar lying experiments, which typically find that people lie surprisingly little or not at all (Abeler et al. (2014), Abeler et al.

A



B

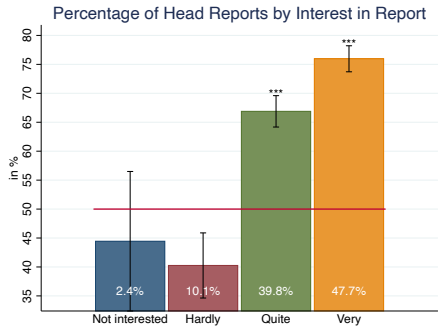


Figure 2.1: **Proportion of mayors who report heads and tails and interest in receiving the report**. (A) The percentage of mayors reporting heads and tails is displayed above the bars, showing that they differ significantly from the objective 50% benchmark (two-sided binomial test), indicating a high frequency of lying. (B) The percentage of mayors who reported heads depending on their interest in the report is given by the height of the bars and additionally, at the bottom of the bars, the share of mayors in each category of interest in the report is displayed. Standard errors around the mean are given by the intervals. Stars indicate a significant deviation from the 50% benchmark calculated by a two-sided binomial test, *** p-value<0.01.

(2019)). While these results appear to confirm the stereotype that politicians are likely to lie, in our game, there was extensive variation in this behavior. We then assess which observable characteristics are associated with preferences for truth-telling among mayors. The evidence suggests that women are equally likely to lie as men, but mayors of large parties lie more often. Importantly, we find that dishonesty is significantly correlated with being reelected in our sample, even when controlling for actually standing for reelection. The finding that dishonest mayors are more likely to survive in office suggests that dishonesty may confer advantages in politics.

This paper uses a behavioral lying game to study lies among politicians. It hereby contributes to the growing literature in political science that uses behavioral games to study the dispositions of political elites, which has so far focused on traits such as the tendency to escalate commitment, status quo bias, and future discounting (Sheffer et al. (2018), for a review, see Hafner-Burton et al. (2013)). We also contribute to the empirical literature in behavioral economics by adapting standard lying games to a population where their administration is not feasible and by focusing on a population that has not been studied before.

2.2 Results

Figure 2.1 shows the frequency of lying in our study. It depicts two key findings. First, a substantial proportion of politicians lied. We find that nearly 68% of subjects reported heads, as shown in Figure 2.1 A. The empirical distribution is significantly different from the expected 50% if everyone was telling the truth, which is confirmed through a two-sided binomial test (p-value$<$0.01). This high frequency of lying differs from that found in the most similar designs in the literature. For instance, Abeler et al. (2014) administered a truth-telling experiment to a general population sample in which they asked respondents to flip a coin four times and provided monetary rewards for obtaining

tails. The distribution of the reported outcomes is indistinguishable from the truthful distribution. In a one-shot game administered to a larger sample, only 44% of the sample reported the winning coin flip outcome (they hypothesize that some people lied to their monetary disadvantage due to privacy or self-image concerns). Another study, Cohn et al. (2014), finds that 52% of a control sample of bankers and 58% of a banker sample framed in terms of their professional identity reported the coin flip outcome that led to a monetary reward.

Second, Figure 2.1 B demonstrates that the nonmonetary incentives used in this research were a powerful motivator for lying. A large majority of mayors was interested in receiving a report of the results of the survey, with 48% reporting that they were very interested and 40% reporting that they were quite interested. The reported outcome of the coin toss varied sharply depending on interest in the report, with 76% of those very interested and 67% of those quite interested reporting heads (both significantly different from 50%, two-sided binomial test p-value$<$0.01), compared to 44.5% and 40% among those who were not at all or only somewhat interested respectively (both not statistically different from 50%, two-sided binomial test p-value$>$0.1). These results indicate that the prospect of receiving the report incentivized lying particularly well among those who valued the reward most, suggesting that our incentivization mechanism is a valid alternative tool to monetary rewards when studying lying among political elites.

2.2.1 Gender

Would honesty in politics increase if there were more female politicians or are women in politics no different from men? A large literature has investigated whether men or women are more likely to lie in different types of behavioral games and has found that women are not always more averse to lying. Gender differences depend on the type of lie and the probability of being discovered. A meta-analysis of sender-receiver games finds that men are more likely to tell lies than women when

they harm or benefit the receiver but there are no differences in the case of Pareto white lies which benefit both (Capraro (2018)). In games that vary the risk of being detected, men are more likely to lie

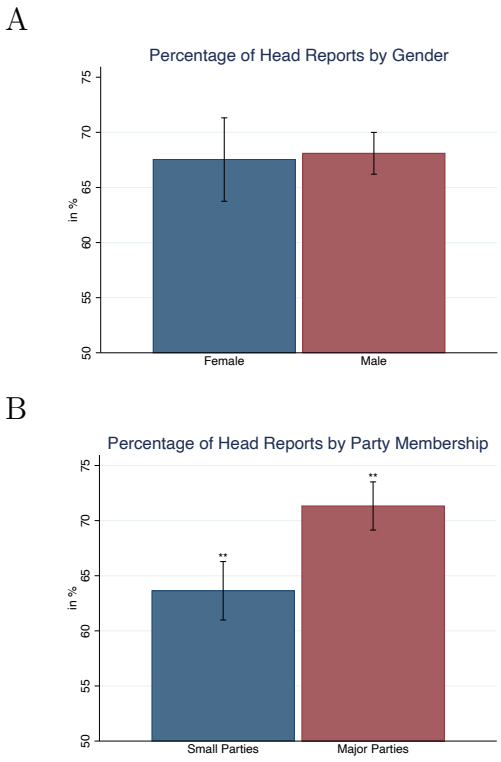


Figure 2.2: **Percentage of mayors who reported heads by their individual characteristics**. (A) The percentage of mayors reporting heads by their gender and (B) membership in a large party. All categories exceed the 50% benchmark (two-sided binomial test). The difference between genders is negligible. However, there is a highly significant difference between members from major versus minor political parties. Standard errors are given as intervals around the mean. Stars indicate significant deviation of reported heads between subgroups (two-sample t test), ** p-value$<$ 0.05

than women when the risk of being detected is high but there is no difference when the risk is low (Kajackaite and Gneezy (2017)).

We are interested in lies that neither directly harm nor benefit others and have no risk of being discovered. These types of lies match situations of interest in politics in which a politician has private information unknown to voters. Our lying game creates such a setting, which should be less conducive to gender differences in lying than settings where other players are directly harmed or the risk of being discovered is high.

Figure 2.2 A confirms that there is no significant gender difference in the percentage of mayors who reported heads. A two-sided binomial test confirms that both the proportions of female and male mayors are significantly different (p-value$<$ 0.001 for both) from the 50% benchmark that we should have observed if people had been truthful on average. A t-test that tests for differences between the percentage of reported heads by male and female mayors supports the null-hypothesis of no difference (p-value=0.89). To assess the robustness of these findings, we conducted a linear probability regression analysis (see Table 2.1 below). As expected, we find that gender does not predict reporting heads in any specification. These results suggest that increasing the number of female politicians, counter to popular stereotypes, would not have a direct impact on the frequency of political lies, at least in the type of situations we study.

2.2.2 Party membership

A key observable characteristic of politicians is their party membership. We examine if politicians from different types of parties differ in their preferences for truth-telling, comparing the two largest nationwide parties (Partido Popular [PP] and Partido Socialista Obrero Español [PSOE]) to the other parties, such as regional and local parties. Large parties may be more likely to contain dishonest politicians for two reasons. First, party membership and dishonesty could be linked via more frequent exposure to dishonest practices in the organizational

structures of major parties. Consistent with this possibility, previous research shows that large parties tend to have larger bureaucratic apparatuses which have been linked to more corrupt structures (Pujas and Rhodes (1999), della Porta (2004)). In the specific case of Spain, scandals revealing systemic corruption have affected the two main parties (Heywood (2007)). Second, dishonesty may be more prevalent in large parties due to self-selection of more dishonest politicians into these parties as their greater access to resources and power provides more opportunities for corruption.

We find that mayors who are members of one of the two major parties in Spain, PP and PSOE, reported heads significantly more than those who are members of smaller parties. Figure 2.2 B shows that 71% of mayors from major parties reported heads, while only 64% of those from smaller parties reported heads. For both groups, a two-sided binomial test rejects the null of truthful behaviour (with p-value$<$0.001 for both) and a t-test for difference in means rejects the null of subgroup equality (p-value=0.02). The regression results (Table 2.1) support this finding. Specifically, the regression coefficient suggests that being a member of one of the major parties increases the chance of reporting heads by eight percentage points. Together with the finding that all groups had significant levels of lying, this suggests that members of major parties lied significantly more.

2.2.3 Reelection

We now turn to the relationship between preferences for truth-telling and political survival. Previous research has shown that politicians who have less agreeable personality traits outperform others on various indicators of political success, including reelection (Joly et al. (2018)). In the case of dishonesty, a correlation could emerge due to two main processes. First, dishonest politicians might be more willing to defy deontological norms in the pursuit of other goals (such goals could be egoistic such as winning office or altruistic such as better representing constituents' interests). If undiscovered or un-

punished, this willingness to defy deontological norms could confer a political advantage at governing and campaigning effectively, resulting in higher political survival. The relationship between honesty and

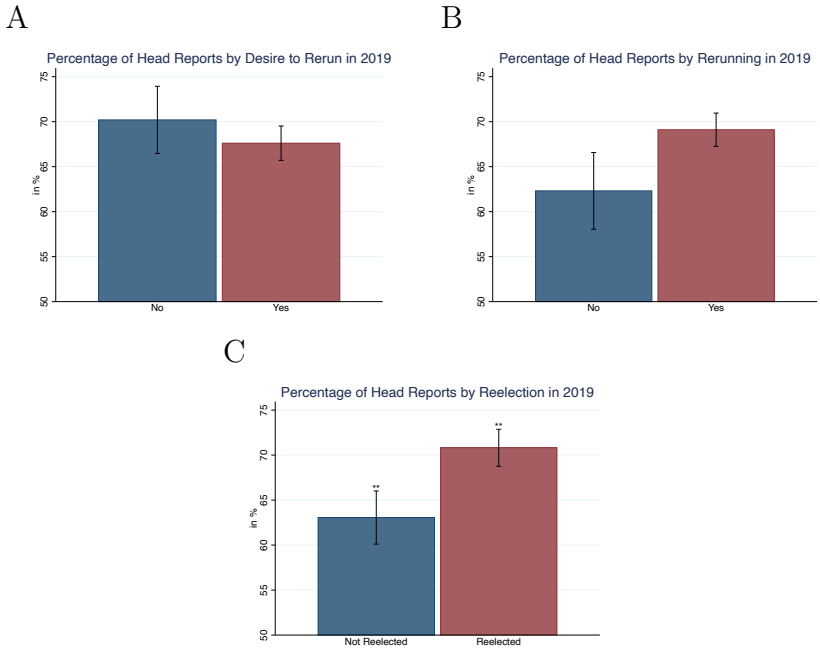


Figure 2.3: **Percentage of mayors who reported heads by measures relevant to reelection**. The percentage of mayors reporting heads by (A) their reported desire to rerun for office, (B) actually rerunning for office, and (C) reelection results. All categories exceed the 50% benchmark (two-sided binomial test). The difference between those who want to rerun and those who do not is small and not statistically significant. Similarly, there is no statistical difference between those who reran for office and those who did not. However, there is a highly significant difference between reelected and not reelected mayors. Standard errors are given as intervals around the mean. Stars indicate significant deviation of reported heads between subgroups (two-sample t-test), ** p-value$<$ 0.05

reelection could also emerge due to self-selection into rerunning for office. Inexperienced, honest politicians should quickly realise that in some situations, governing might be difficult without getting "dirty hands" (Walzer (1973)). They may resent being confronted with such moral dilemmas and decline to run again.

In this section, we study if honest and dishonest politicians differ in their stated desire to rerun for office, in whether they actually compete again, and in their reelection rates. Our survey asked mayors if they would be willing to rerun for office in the next municipal elections, which took place in May 2019, 5 mo after the end of the fieldwork. We also collected data on whether they actually did run again in those elections by examining if they were in the first three positions of the ballot (Spain uses a closed-list PR system, and we find that a nonnegligible number was placed towards the end of the ballot, signaling support for their party but unwillingness to serve as a mayor again. See Materials and Methods.). The stated desire to rerun for office, as well as actually rerunning, measure self-selection and are key control variables in our reelection analysis.

A large majority of mayors sought reelection. In the survey, 80% of mayors reported that they would surely or probably want to run for reelection and the percentage of those actually running for reelection is close to 83%. Specifically, ca. 80% of mayors who reported tails and 84% of those who reported heads reran for office. Importantly, this difference is not statistically significant (t-test, p-value=0.13) so that self-selection effects into rerunning should not be of particular concern. Figures 2.3A and B present the percentage of mayors reporting heads in the lying experiment depending on the stated desire to rerun and on the actual decision. Interestingly, while those who reported no desire to run again chose heads more frequently than those who reported wanting to rerun (70 and 68%), those who subsequently did not run for reelection chose heads less frequently than those who did run again (62 and 69%). One explanation of this discrepancy is that dishonest mayors are more likely to misreport their willingness to run. Consistent with this, we find that of those who reported no desire to

run again but actually did rerun (roughly 8% of the sample), 78.5% reported heads in the lying experiment. This percentage is much lower (60.5%) among those who did not rerun despite responding that they wanted to (6% of the sample), as well as among those who followed through on their stated desire to seek or not to seek reelection (68 and 63.5%, respectively, reported heads). These results are consistent with the claim that dishonest mayors are more likely to conceal their desire to seek reelection.

We now turn to the question of whether dishonest mayors are more likely to survive in office. From the mayors in our sample, 65% were sworn in as mayors again in 2019. Panel C of Figure 2.3 shows that reelected mayors reported heads significantly more than mayors who were not reelected (71% compared to 63%). It is highly unlikely that the behavior of the reelected and that of the not reelected mayors stems from the same distribution (t-test, p-value=0.03). Next, we examine the relationship between dishonesty and reelection success in a regression framework. Specifically, we are concerned that the correlation may not imply that dishonesty facilitates political survival if it is entirely driven by self-selection of dishonest mayors into running for office or if dishonest mayors choose to run in different environments, particularly in less competitive elections, which in turn facilitate reelection. This would be especially problematic if less competitive environments increase the reelection chances of dishonest mayors more than of honest ones. We measure competitiveness as the margin of the seat share that the party with the most seats holds above that of the party with the second most seats obtained in the 2015 mayoral elections (high margins indicate low competitiveness).

Table 2.2 shows the results of the regression models. Reporting heads is associated with an eight percentage point higher likelihood of being reelected (model 1). We find a substantial and significant relationship between reporting heads and reelection even controlling for actually running for reelection in the 2019 municipal elections, the competitiveness of the 2015 election results, log population size,

gender, and party membership (model 2). The relationship is also robust to additionally controlling for the potential interaction between dishonesty and competitiveness (model 3). Importantly, while competitiveness has a large and significant impact on reelection success, we do not find a differential effect of competitiveness on dishonest compared to honest mayors. Restricting the sample to those mayors who reran for reelection yields similar results (model 4). These results are consistent with the claim that dishonesty confers an advantage for political survival that goes beyond differences in self-selection and in the competitiveness of elections.

The result that dishonest mayors are more likely to be reelected can provide a microfoundation to the well-known finding that discovered corruption is often not, or only mildly, punished electorally (Ferraz and Finan (2008), Mondak (1995), De Vries and Solaz (2017)). Our findings suggest that undiscovered lying promotes electoral success, perhaps because it allows politicians to gain an advantage over their opponents while avoiding the possible electoral costs. It seems likely that a tendency to lie increases the likelihood of engaging in corrupt behavior as well. If corruption scandals erode support for a politician, this could offset the advantage conferred by other undiscovered dishonest behavior. Ultimately, the two effects would cancel each other out and lead to the finding of no correlation between corruption scandals and reelection success.

2.3 Discussion

Our paper introduces a version of the standard coin flip honesty experiment that is well suited to study preferences for truth-telling among politicians. Rather than incentivizing lying through monetary incentives, as is standard in lying games, we used a report of the survey results as an incentive. This modification allowed us to bypass politicians' aversion to monetary compensation because of concerns that receiving compensation may be perceived as engaging

in corruption. This nonmonetary measure is very successful at incentivizing lying, as a large and significant percentage of mayors lied in our study. Our procedure measures lying aversion in a type of setting where the lie is not observable and only directly affects the liar. This is representative of situations in which politicians have access to private information that they can misreport or manipulate to their own advantage, thereby reducing the ability of voters to hold them accountable. Such situations are common in politics, but further research is needed to study behavior in other situations such as when politicians' lies have a clear negative impact on someone else or are easily detectable.

Using the modified lying game, we first discover that some observable characteristics are predictive of lying behavior among politicians. While we find no gender differences in lying behavior, members of major parties are more likely to lie than others. While we do not claim that one can simply look at individual attributes and detect dishonesty among politicians, our results do suggest that there is systematic variation in lying aversion.

The finding that dishonest mayors are more likely to be reelected in the next municipal elections, and that this relationship is not solely driven by differences in self-selection or the competitiveness of the environment, is consistent with the possibility that being dishonest confers some political advantage and facilitates survival. Such advantage could stem from two different mechanisms. First, it could be related to a different policy-making style if politicians are willing to be dishonest in order to achieve their, or their constituents', goals and the achievement of such goals is then rewarded by voters. Second, even if they do not differ in terms of policy-making, dishonest politicians who are willing to distort the truth might communicate and campaign more effectively, resulting in higher popularity and reelection rates. These two possibilities have very different normative implications, ranging from the interpretation that occasionally suspending conventional moral norms can improve the ability to achieve policy goals to the interpretation that dishonesty constitutes an added obstacle

to political accountability. Therefore, it is relevant to first assess the generalizability of these results by replicating the key findings in other settings and to identify the underlying mechanisms that may link dishonesty to political survival.

2.4 Materials and Methods

2.4.1 Data Availability

The data and code files to replicate the results of the paper have been deposited at the Harvard Dataverse and are available at https://doi.org/10.7910/DVN/MPAZUD (Janezic and Gallego (Deposited June 7 2020)).

2.4.2 Setting, Participants, and Fieldwork

In order to study honesty among politicians, we fielded an original survey administered to Spanish mayors. Spain is an excellent setting for our study. It is an advanced industrial democracy which ranks 13th out of 27 European countries in a combined quality of government score (Charron et al. (2014)). In this sense, Spanish mayors are more typical of the population of interest than, for example, Scandinavian politicians, who have received extensive attention because of the abundance and quality of data in countries like Sweden (Dal Bó et al. (2017), Besley et al. (2017)) but who might be outliers in a comparative perspective. The Spanish local institutional setting and the capacity of municipal governments are also fairly typical for advanced industrial democracies. The political system is decentralized with elected governments at the national, regional, and municipal levels. Municipal spending amounted to 14% of total public expenditure in 2007 according to the Organization for Economic Co-operation and Development, a figure similar to countries like Germany, Austria, and Portugal. Local councilors are elected every 4 y with the number of councilors depending on population size. In municipalities with more

than 250 inhabitants, citizens elect councilors using a closed party list proportional representation system. Councilors then elect a mayor, who is the head of the party list which has obtained an absolute majority of votes in the investiture vote. If no party commands an absolute majority of votes from councilors, the head of the party list with the most votes from voters in the municipal elections becomes mayor. In practice, this implies that in more than 90% of cases the head of the most voted party list also becomes the mayor.

Our questionnaire included a range of questions about mayors' background experiences, outside options, political ambition, and political preferences, as well as an embedded lying aversion measure. The survey was programmed and administered online and was pretested through cognitive interviews with 12 politicians who were not in our sample and adjusted according to the feedback received. Survey invitations were sent to all 2,282 Spanish municipalities with more than 2,000 inhabitants. We collected the official email addresses of mayors by consulting websites and calling the municipalities. Informed consent was obtained from all participants. The consent form provided accurate information about the goals of the study, the data handling procedures, the relevant legislation, and the contact details of the principal investigator. Ethical approval was not required, but all materials were reviewed by a legal advisor. In July 2018, we launched a pilot study with mailings to two autonomous communities. We made further adjustments to the questionnaire based on an analysis of the initial 80 responses and the feedback received from participants by email. The main fieldwork was conducted between September 2018 and January 2019.

In order to maximize control over data collection, we did not subcontract the fieldwork to a survey company but conducted it in-house by hiring and training research assistants. We sent up to five reminders by email. In addition to emails, we made phone calls to all municipalities that had not responded. We tried to talk with the mayor, or their secretaries if that was not possible, and sent personalized invitation emails after these conversations. An important

concern was that mayors may delegate responding to surveys to subordinates. To address this issue, we sent the invitations to the official email addresses of mayors rather than to generic institutional addresses, and we stressed in the invitation email and on the first page of the survey that the survey had to be taken by the mayors themselves. We cannot rule out that in a few cases the survey was filled in by an aide of the mayor, yet examining responses to open-ended questions about personal information unlikely to be known by aides suggests that the survey was answered personally by the vast majority of mayors. Specifically, we find that only 31 out of 816 respondents did not fill in the occupations of the mayors' fathers and mothers, respectively. This is a very high response rate to an open-ended question that an aide answering the survey would be very likely to skip since answering the question was nonobligatory. Participation in the survey was not compensated. In the pilot study, we embedded a financial reward in a lying aversion game, but as explained below, we decided to eliminate any monetary compensation due to the strong complaints it generated. The simplified version of the honesty game was not preregistered. Further details about the consent form and the questionnaire can be found in SI Appendix. We collected a total of 816 full responses to the lying aversion measure, which represent 36% of the population, an average to high response rate in elite surveys.

2.4.3 New measure of lying

Self-reported integrity is not a trustworthy measure of honesty and may even be negatively correlated with actual honesty. To circumvent this difficulty, previous research has measured lying aversion through a variety of behavioral methods (for reviews, see Rosenbaum et al. (2014) and Jacobsen et al. (2018)). In our setting, observability of the decision to lie is particularly relevant because politicians are concerned about maintaining a reputation for honesty. We chose a nonobservable task to reduce the risk that politicians' decisions

were motivated by the wish to appear honest to us. Such tasks were introduced by Fischbacher and Föllmi-Heusi (2013), who ask subjects to roll a die privately and to report the outcome, with higher values leading to higher payoffs. To identify the share of liars, researchers compare the theoretical distribution of how often each die side should come up with the distribution of reported outcomes. An alternative version of this experiment uses a coin rather than a die (Abeler et al. (2014), Cohn et al. (2014)). Because lying is incentivized equally for all individuals, and reputational concerns are eliminated when individual decisions are not observable, differences in the prevalence of lies can be attributed to lying aversion. Subgroup analysis can reveal which individual characteristics are associated with lying aversion. We explicitly told mayors that we cannot observe the true outcome of their action. Due to the nature of the population of interest, we needed a task that was quick and easy to conduct and therefore used a coin flipping rather than a die rolling task. In case some of the mayors did not have a coin at hand, we provided a link to a website that virtually throws fair coins.

In lying aversion experiments, it is customary to use money as an incentive because subjects are assumed to derive a similar level of utility from any given amount of money. However, offering monetary incentives was not feasible in this population. Politicians are subject to more scrutiny than other citizens and are often accused of benefiting economically from holding office. Thus, we suspected that politicians might feel more uncomfortable than other populations when being offered a monetary reward. Yet, to keep with the literature, we designed a lying game with monetary incentives (in line with legal limits on what politicians are allowed to accept) and included it in a pilot study that was answered by mayors. However, we received multiple emails and phone calls from mayors who were offended by our attempt to pay them a monetary reward. Considering that mayors are generally very busy, the fact that they took the time to complain about an academic study should further illustrate how big a problem using monetary rewards presents. Instead of serving as an incentive,

money served as a disincentive.

The pilot included an alternative, nonmonetary lying aversion measure, and in this case we found that the incentive was very well received. We decided to use the desire to know how mayors compare to other politicians as the incentive device for the lying experiment. At the beginning of the survey, we asked mayors whether, and how strongly, they would like to receive a personalized report about the survey results. As shown in Figure 2.1 B, mayors were interested in receiving the report. At the end of the survey, we told them that we could only send the reports to some mayors. A coin flip that the mayors themselves had to conduct and then report decided whether they would receive it or not. If they reported heads, they received the report, and if they reported tails, they did not. If mayors had an interest in receiving the report, then they had an incentive to report heads irrespective of the actual outcome and thus to lie. The expectation was therefore that compared to the theoretical distribution, many more head reports would occur. This is confirmed by our results discussed in Results. Our identifying assumption is that mayors do not lie to their disadvantage.

2.4.4 Statistical Analyses and Regression Model

In order to estimate the relationship between personal characteristics and dishonesty, we conducted linear probability regressions of the following form:

$$P(Heads) = \alpha + \beta_1 X_1 + \beta_2 Int.Rep. + \beta_3 Controls + \varepsilon \qquad (2.1)$$

We control for interest in the report by including a dummy that takes value 1 if the mayor reported that he or she was interested or very interested in the report. For the gender and party analyses, we use the probability to report heads as the dependent variable and gender and membership in a major party as the independent variables (X_1) respectively. The coefficient estimates can be used directly to compute the effect of the independent variable. For example, the estimated

coefficient of major party is 0.08, which implies that being a member of a major party increases the probability of reporting heads by eight percentage points.

Importantly, all our estimates are likely to be a lower bound. This is because the group of mayors who reported heads includes both dishonest mayors who obtained tails and lied and honest mayors who obtained heads and reported truthfully. This latter group should be similar to honest mayors who obtained tails and reported tails. The mixed composition of our heads group implies that our estimates should be larger if we could isolate dishonest mayors.

For the reelection analysis, we use reelection as the dependent variable and reported heads as the independent variable to model the claim that political outcomes are affected by politicians' honesty rather than the other way around. The linear probability regression equation is

$$P(Reelected) = \alpha + \beta_1 Heads + \beta_2 Rerunning + \beta_3 Controls + \varepsilon \tag{2.2}$$

We control for the effects of the other variables of interest to ensure that the large relationship between reelection and reporting heads is not driven by one of the other characteristics such as party membership. Here we control for party membership by using party dummies rather than a dummy for membership in one of the largest parties as reelection results are impacted by the party itself rather than the size of the party. As a supplementary analysis, we restrict the sample to those mayors who reran for election and show that the relationship between dishonesty and reelection holds for this subsample.

For both specifications, we added controls such as log population size, the margin of the percentage of seats held by the party with most seats in the council compared to that of the party with the second most seats, and mayors' ages. We find that the relationship is robust to such controls. As some of the additional controls are not available for all mayors, our regressions contain slightly fewer observations than the 816 full responses. In addition, we excluded the five mayors who

did not respond to the question about their interest in the report.

To assess whether behavior is significantly different from the theoretical 50% benchmark, we conducted binomial tests. We used unpaired two-sample t-tests with a two-sided alternative to test whether the mean behavior of mayors in the categories into which we subdivided them, e.g. ,main party members versus minor party members, is significantly different from each other.

We excluded the 10 mayors from the analysis who took less than 5 s to fill out the coin flipping question. As it was impossible to read the question, take out a coin, flip it and report the result within 5 s, these mayors lied to us about having performed the coin flip at all. To prevent potential bias from introducing an additional lie element to the data, we excluded them. In addition, we excluded mayors who took more than 90 s to complete the question as it is likely that their responses are of lower quality due to inattention. For robustness, we check if this influences the results and find that they are robust to including these mayors (SI Appendix).

2.4.5 Robustness Analyses

To assess whether our sample is representative of the Spanish mayor population, we compare our sample to the whole population of mayors in relation to the share of women, average population size, percentage with a university degree, mean age, mean turnout, and the shares of major parties and national parties. We find that our sample is representative of the population for most characteristics (SI Appendix, Table S1) but that mayors are on average two years younger and that municipalities are smaller in our sample. The size of the difference regarding age is small enough that we do not have to be concerned with selection issues. For municipality size, it is important to note that the difference is largely driven by the size of the largest municipalities in Spain. We also conduct an out-of-sample prediction analysis to test whether our sample is biased and find that there is no statistically significant difference in predicted behavior compared

to actual behavior (SI Appendix, Table S2).

We next examined whether our results are robust with respect to removing mayors who took a long time answering the coin flip question. To that end, we reran the regressions of interest using cutoffs with 5 s less or more than the cutoff of 90 s, as well as removing the cutoff entirely. The results (SI Appendix, Tables S3 and S4) are robust in terms of the sign, size, and significance. We also examine whether our results are robust with respect to the standard errors that we chose. Therefore, we reran all analyses with heteroskedasticity robust standard errors. We find that our results are robust to changing the standard error specification (SI Appendix, Tables S5 and S6). A possible issue with using linear probability models is that the fitted values might not be bounded between 0 and 100%. We examine the fitted values (SI Appendix, Figure S1) and find that the majority lies between 60% and 90%. This means that the theoretical unboundedness of linear probability models is not an issue.

Acknowledgements: We thank the editor and anonymous reviewers, Larbi Alaoui, Jose Apesteguia, Andreu Arenas, Antonio Cabrales, Andrew Clark, Ray Duch, Andy Eggers, Martín Fernández-Sánchez, Lukas Hoesch, Gerda Hooijer, Gaël Le Mens, Rosemarie Nagel, Nikolas Schöll and participants at workshops and conferences at International Society of Political Psychology 2019, Universitat Pompeu Fabra, Université Catholique Louvain la Neuve, Uppsala University and Paris School of Economics for their comments as well as Alba Huidobro, Yeimy Ospina and Nicolas Bicchi for their research assistance. We gratefully acknowledge financial support from the Spanish Ministry for the Economy, Industry and Competitiveness (grant CSO2016-79569-P).

Table 2.1: Linear probability regressions with gender and membership in a major party as independent variables.

	(1)	(2)	(3)	(4)	(5)
	Rep. Heads	Rep. Heads	Rep. Heads	Rep. Heads	Rep. Heads
Interest Report	0.31***	0.33***	0.31***	0.33***	0.33***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Gender	0.00	0.00			-0.00
	(0.04)	(0.04)			(0.04)
Major Party			0.08**	0.08**	0.08**
			(0.03)	(0.03)	(0.03)
Population size, log		-0.02		-0.02	-0.02
		(0.02)		(0.02)	(0.02)
Age		-0.00		-0.00	-0.00
		(0.00)		(0.00)	(0.00)
Margin 2015		-0.12		-0.14	-0.14
		(0.11)		(0.11)	(0.11)
Constant	0.41***	0.67***	0.37***	0.63***	0.63***
	(0.05)	(0.19)	(0.05)	(0.19)	(0.19)
Observations	759	700	759	700	700

Standard errors are in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

The dependent variable is a dummy for whether a mayor reported heads.

Table 2.2: Linear probability regressions with reelection as dependent variable.

	(1) Reelected	(2) Reelected	(3) Reelected	(4) Reelected
Reported Heads	0.08**	0.05*	0.08*	0.10*
	(0.04)	(0.03)	(0.05)	(0.06)
Ran for Reelection		0.77***	0.77***	
		(0.04)	(0.04)	
Margin 2015		0.40***	0.52***	0.66***
		(0.09)	(0.15)	(0.19)
Gender		-0.02	-0.02	-0.04
		(0.03)	(0.03)	(0.04)
Population size, log		-0.00	-0.00	-0.00
		(0.01)	(0.01)	(0.02)
Reported Heads × Margin 2015			-0.17	-0.26
			(0.18)	(0.22)
Constant	0.59***	0.01	-0.02	0.72***
	(0.03)	(0.14)	(0.14)	(0.17)
Party Dummies	No	Yes	Yes	Yes
Observations	758	754	754	624

Standard errors are in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

The dependent variable is a dummy for whether a mayor was reelected.

Model 4 reports the results for the sample restricted to mayors who reran for election.

Chapter 3

REASONING ABOUT OTHERS' REASONING

(joint with Larbi Alaoui and Antonio Penta)

This paper has been published in the Journal of Economic Theory in 2020, see the full citation below:

Alaoui, L., Janezic, K. A., & Penta, A. (2020). Reasoning about others' reasoning. Journal of Economic Theory, 189, 105091. Chicago

URL: https://www.sciencedirect.com/science/article/pii/S0022053120300855?via\%3Dihub

DOI: https://doi.org/10.1016/j.jet.2020.105091

3.1 Introduction

Recent experiments have documented that, in games in which individuals behave according to standard models of level-k reasoning, changing subjects' beliefs affects the observed distribution of levels (see, e.g., Agranov et al. (2012), Georganas et al. (2015) and Alaoui and Penta (2016a)). These results suggest that, at least in some settings, level-k patterns of behavior may be driven by individuals' beliefs rather than by their intrinsic cognitive limitations. Whether this is true in general, however, is far less clear, and most work in this area is agnostic on the point. This is largely because disentangling 'behavioral' and 'cognitive' levels in the lab can be difficult, and in fact even theoretical models do not typically distinguish the two.[1]

One exception is provided by the Endogenous Depth of Reasoning (EDR) model of Alaoui and Penta (2016a) in which a subject's understanding of a game (his *cognitive bound*, or *capacity*) is formally distinct from his 'behavioral level'.[2] In the EDR model, holding constant an individual's cognitive bound, the observed level of play may vary with the individual's beliefs about the opponents. For instance, even if a subject understands up to five iterations of the level-k reasoning, he may sometimes play as a level-5, but he may instead play as a level-3 if he thinks the opponent would play as a level-2. The EDR model also allows players' very understanding, or capacity, to vary with the stakes of the game. For example, a subject may understand three iterations of the reasoning process when the stakes are low, but more when the stakes are high enough, depending on his cognitive

[1] Friedenberg et al. (2017) address a similar question, but from a different perspective. We discuss the differences and similarities with their work in Section 3.1.1.

[2] Alaoui and Penta (2016a) use the term 'cognitive bound' to refer to the highest level-k that a subject is able to conceive of, in a given game, which indirectly provides an upper bound to his behavioral level in the game. The term 'capacity' is due to Georganas et al. (2015), essentially with the same meaning. Friedenberg et al. (2017) also use the term 'cognitive bound', but with a different meaning (see Section 3.1.1).

abilities.

The EDR model provides a formal language with which to ask whether in practice level-k patterns of behavior are driven by subjects' cognitive bounds or by their beliefs, possibly of higher order. But the experimental treatments in Alaoui and Penta (2016a) (AP hereafter), which vary subjects' beliefs as well as the stakes for all players at the same time, shed little light on this particular question. AP's treatments also do not disentangle the extent to which the more sophisticated behavior observed in the 'high stakes' treatments is due to agents' deeper understanding or to their beliefs about the increased depth of reasoning of their opponents. Conceptually, the two points are related: if subjects did not reason at all about others' incentives to reason, then the higher sophistication in the 'high stakes' treatments would be exclusively driven by their own increased capacity; if, instead, behavior were purely determined by beliefs, then subjects' behavior should not change, if nothing changes about the opponents. The challenge is to identify these effects in the lab, which is the objective of the present paper.

To assess the extent to which agents' behavior is determined by a binding cognitive bound, as opposed to beliefs, we design treatments based on the following simple idea, which we call the *tutorial method*. Suppose players are engaged in a standard game for level-k reasoning, such as a version of Nagel (1995)'s beauty contest.[3] Now entertain the following thought experiment: take a subject, say Ann, whose choice in this game is consistent with level-3 behavior, and provide her with a game theory tutorial which explains the strategic structure of the game (best responses, iterated reasoning, uniqueness of equilibrium, etc.), but without providing any proper factual information (such

[3]This paper will focus on another game, referred to as the *acyclical 11-20 game*. The logic of the argument would be the same, but we use Nagel (1995)'s beauty contest game here because it has been one of the main workhorses for level-k reasoning (see also Camerer (2003), Crawford et al. (2013) and references therein). Other prominent games in the level-k literature include the two-person guessing games of Costa-Gomes and Crawford (2006) and Basu (1991)'s travelers' dilemma (e.g., Capra et al. (1999)).

as information about others' choices, typical distributions of actions in this game, etc.). Next, ask Ann to play this game again, but against individuals who have *not* received the tutorial. Intuitively – setting aside difficulties in computing the best responses, noise in Ann's reasoning or choice, and other caveats – if Ann perceives the new pool of opponents as identical to those in her first trial, then her action should change only if the game theory tutorial has made her understand something she deems useful. So, if her level-3 action in the pre-tutorial treatment was purely driven by her beliefs about the opponents (e.g., that they behave as level-2's), then the tutorial should have no impact on her choice, and her behavior would be level-3 in both rounds. However, if her action shifts (and especially if it shifts towards a higher level-k), then it must be the case that her previous understanding was in some sense 'binding', and hence her level-3 choice was not entirely due to her beliefs.

Our second question – understanding whether agents explicitly take into account others' incentives to reason – is more directly motivated by the central premise of the EDR model, which is that agents' cognitive bound may itself vary with the payoffs of the game. As explained above, however, the point is inherently related to the broader problem of the cognition-beliefs dichotomy. But disentangling own understanding from reasoning about others' incentives to reason presents non-trivial conceptual difficulties. For instance, suppose – as assumed in the EDR model – that subjects' cognitive bounds are increasing in their own stakes in the game. Then, intuitively, one way to disentangle the two effects is to consider 'asymmetric transformations' of payoffs in a two-player game, in which stakes are increased for one player (Ann) but not for the other (Bob). This change, however, would not be enough to isolate the effects on Ann's own understanding, because she may think that Bob could react to her stronger incentives to reason. If this were the case, then a change in Ann's behavior need not be driven by her own understanding, but by her beliefs about Bob's reaction to her incentives. In other words, to isolate the effects of Ann's higher stakes on her own understanding,

it is important to hold constant Ann's beliefs about Bob's reasoning, of *any* order. For this reason, we design treatments to disentangle own reasoning from reasoning about the opponents' reasoning. These treatments make use of what we call the *replacement method*. That is, in the asymmetric payoff treatment, Ann does not just play against a subject whose stakes are low; rather, Ann plays against the choice made by a player, Bob, who is engaged in a game in which stakes are low for both players (hence, Bob's opponent is not Ann: in our treatment, Ann is 'replaced' by a low-stakes version of herself). This way, Ann's beliefs (of any order) about Bob's reasoning are identical to her beliefs in the low-stakes game, and hence any change in behavior observed when Ann's stakes are increased can be unambiguously imputed to Ann's own incentives.

We apply both the *tutorial* and the *replacement* methods to three experiments. Experiment 1 leverages the existing dataset by applying these methods to the baseline experiments in Alaoui and Penta (2016a). Experiments 2 and 3 instead develop simpler variations of those treatments for a new pool of subjects, and address possible concerns on demand effects associated with the specific application of the tutorial method. A fourth experiment addresses possible robustness concerns on social effects which may arise in some of the treatments of Experiment 1. All experiments are based on the 'acyclical 11-20 game' from Alaoui and Penta (2016a), but the underlying logic has broader validity and does not rely on any specific feature of the game nor of the EDR model. Hence, our experimental designs suggest a general methodology that can be easily extended to other games and settings: The *tutorial method* can be used to investigate the cognition-beliefs dichotomy in general models of strategic thinking, and the *replacement method* can be used to explore higher-order beliefs effects in general games, with essentially no restrictions on the underlying payoffs.

Methodological considerations aside, our empirical findings show that, for a large fraction of subjects, the cognitive bound is actually binding when they play against opponents who are regarded as more sophisticated. This is a perhaps surprising result for the view that

level-k behavior is mainly driven by beliefs: it suggests that, at least in some settings, level-k models are directly applicable to agents' own understanding. On the other hand, we also find evidence that a large fraction of subjects do reason about others' incentives to reason, providing support to a much more subtle implication of the EDR model than those that were previously tested. Overall, our results suggest that level-k behavior in general should not be taken as driven either by cognitive limits alone or beliefs alone: it depends on the complex interaction of cognitive bounds, beliefs about opponents' cognitive abilities, and reasoning about the opponents' reasoning processes. We also find that the EDR framework is a useful tool for analyzing and understanding this interaction, and that the results are overall consistent with its predictions.

The rest of the paper is organized as follows: Section 3.1.1 reviews the related literature, Section 3.2 introduces the baseline game and logistics common to all experiments, and Section 3.3 presents the specific treatments. Section 3.4 contains theoretical results on the EDR model which are relevant for the treatments in our experiment, and spells out the identification assumptions used to connect the model to the experimental findings. Section 3.5 presents the experimental results. Section 3.6 concludes.

3.1.1 Related Literature

The classical literature on the level-k and cognitive hierarchy models (e.g. Nagel (1995); Stahl and Wison (1995); Costa-Gomes et al. (2001); Camerer et al. (2004); Costa-Gomes and Crawford (2006)) has analyzed systematic features of observed behavior which suggested that individuals follow distinct patterns of reasoning. This evidence has often being interpreted as being driven by individuals' limited ability to reason strategically, but models in this literature are typically silent on whether the observed 'levels of play' stem from subjects' cognitive limitations, or perhaps from their beliefs about others' rationality (of any order) or their ability: most models are consistent with both

interpretations.[4]

More recent experiments have focused on how levels of play vary across different games and with different opponents (e.g., Agranov et al. (2012); Georganas et al. (2015) and Alaoui and Penta (2016a)). Their findings suggest that, at least in some settings, level-k patterns of behavior may be driven by individuals' beliefs rather than by intrinsic cognitive limitations. The distinction between 'cognitive' and 'behavioral' levels – that is, between the maximum level-k an agent can conceive of, due to his limited ability, and the level of his action, which may be driven by his beliefs – has been made explicit in some recent theoretical models: for instance, Strzalecki (2014)'s notion of level-k type only restricts the support of a type's beliefs, but level-k behavior may vary as a type's beliefs are varied; similarly, Alaoui and Penta (2016a) define the 'cognitive bound' as the maximum level an agent can conceive of, but that's distinct from the 'behavioral level', which is jointly determined by the cognitive bound and the agent's beliefs; Georganas et al. (2015) also have an analogous distinction, and use the term 'capacity' essentially with the same meaning as Alaoui and Penta (2016a)'s 'cognitive bound'. Similar ideas have been extended to dynamic games by Rampal (2018a). Rampal (2018b) also finds evidence of behavior driven by agents' beliefs.

Friedenberg et al. (2017) study a related problem concerning a rationality-cognition dichotomy, but where cognition refers to a distinct concept from ours. More specifically, in Alaoui and Penta (2016a) and Georganas et al. (2015), the cognitive bound or capacity refers to a player's understanding of the game in the sense of the level-k literature, that is as the highest level of iteration of best replies the player is able to conceive of (though, as we discussed, not necessarily the one he plays). In contrast, Friedenberg et al. (2017) depart from the level-k literature in that they define a player to

[4]This literature has also spurred more sophisticated game theoretic work, which has tackled the challenging question of how to model players who can only conceive of finitely many orders of beliefs, and studied the behavioral implications of those situations (see, e.g., Kets (2017) and Heifetz and Kets (2018)).

be 'cognitive' if his behavior responds – in any way, rationally or not – to changes in payoffs. This provides a measure of cognitive bound, which in their analysis identifies a lower bound to individuals' reasoning ability. Similar to ours, their measure of cognitive bound is also at least as large as their rationality bound (which in turn is conceptually analogous to our behavioral level, although formally distinct), but it may be strictly larger than the cognitive level in our sense. Applying this broader notion of cognition to the experimental data from Kneeland (2015), they find evidence of a significant gap between subjects' rationality and cognitive bounds, and hence of their reasoning ability.

In AP's EDR model the cognitive bound is endogenously determined by a player's cognitive abilities (represented by costs of reasoning) and the incentives to reason (which depend on the game's payoffs). AP test the main predictions of the EDR model with the baseline treatments in Section 3.3.1.1, and show how it can be used to perform robust predictions across games as well as explain the experimental findings in Goeree and Holt (2001)'s famous 'little treasures' experiments. Alaoui and Penta (2018) provide an axiomatic foundation of the model, by characterizing the properties of the reasoning process that justify a cost-benefit approach, as well as more special functional forms for the value of reasoning. Recent extensions of the approach include Alaoui and Penta (2016b; 2018), which extends the EDR model to account for response time, with an application to the experiment by Avoyan and Schotter (2020).

Gill and Prowse (2016; 2017) also investigate more explicitly the connection between level-k behavior and cognitive or non-cognitive abilities, but in a setting with feedback, thereby focusing on learning. They find significant effects of different IQs on the speed of learning, but not on the initial responses.

3.2 Baseline Game and General Logistics

The experiments are designed not only to test whether individuals play differently when their incentives and beliefs about opponents change, but also to analyze the direction in which their actions change, i.e., towards higher or lower level-k's. Moreover, we aim to disentangle whether their action is dictated by their cognitive constraints, given their incentives, or by their beliefs over their opponents' cognitive constraints. These objectives are reflected in the choice of the baseline game, in the logistics of the experiment and in the subject's classification criteria. In this section we discuss each of these elements of our design.

3.2.1 The acyclical 11-20 game

The baseline game remains the *acyclical 11-20 game* throughout:

> The subjects are matched in pairs. Each subject enters an (integer) number between 11 and 20, and always receives that amount in tokens. If he chooses *exactly one less* than his opponent, then he receives an extra x tokens, where $x \geq 20$. If they both choose the same number, then they both receive an extra 10 tokens.

This game is a variation of Arad and Rubinstein (2012)'s '11-20' game. The only difference is that the original version does not include the extra reward in case of a tie. As argued by Arad and Rubinstein, the 11-20 game presents a number of advantages in the study of level-k reasoning, which are inherited by our modified version (see Alaoui and Penta (2016a)):

First, level-k reasoning is the obvious focal way in which to approach the game. This is useful because our goal is to examine the effects of changing beliefs or payoffs on the distribution of levels rather

than to assess the effectiveness of level-k models relative to competing methods of reasoning.

Second, best responding to any level of reasoning is straightforward. A level-1 player chooses 19, the level-2 best response is to play 18, level-3's best response is to choose 17, and so on. The simplicity of the set of best responses is desirable as we do not seek to capture cognitive limitations stemming from computational complexity.

Third, playing 20 is a natural starting point for the iterative reasoning process. Furthermore, it is the optimal choice for any player who disregards all strategic concerns. This level-0 specification is thus intuitive and straightforward.

Fourth, the reasoning process is robust to the possibility that multiple level-0 strategies exist as playing 19 is the level-1 best response for an extensive list of level-0's, such as choosing 20, or the uniform distribution over the action space.

In addition to these points, our modification of the game leads to another useful feature for our objectives. By introducing the extra reward in case of a tie, the best response to 11 is 11, and not 20, as in the version of Arad and Rubinstein. Thus, our modification breaks the cycle in the chain of best responses, which enables us to assign one specific level of reasoning to each possible announcement (with the exception of 11, which corresponds to any level equal to 9 or higher): Action 19 can only be a level-1 strategy, 18 can only be a level-2 strategy, and so forth for every k up to $k = 8$.[5] In the

[5]The fact that every action in this game corresponds to some level-k makes the 11-20 unfit to test level-k models against alternative models of reasoning. That objective would be better attained considering games with large strategy spaces, such as Nagel (1995)'s original beauty contest or Costa-Gomes and Crawford (2006)'s two-person guessing game. As explained above, however, our objective is to test properties of level-k reasoning when beliefs and payoffs are varied, not to contrast level-k with alternative theories. The latter problem has been the focus of an already extensive literature, which overall has provided strong support for level-k reasoning in a variety of settings. For recent work in this direction, see Kneeland (2015). For a more nuanced view, Goeree et al. (2017) argue that a version of level-k models with noise (namely, the noisy introspection model, see

original 11-20 game, action 19 could have been played by a level-1, but also by a level-11, level-21, or other 'high' levels (levels of form $10n + 1$). Although levels-11 and above appear to be uncommon, it is crucial that these cycles be avoided here. That is because some of the hypotheses that we aim to test concern shifts in the distribution of level-k's, but these hypotheses could not be falsified in the presence of such cycles.

3.2.2 General Logistics

The subjects of all experiments are undergraduate students from different departments at the Universitat Pompeu Fabra (UPF), in Barcelona. There were 278 subjects in total, with 120 participating in Experiment 1, 60 in Experiment 2, and 34 in Experiment 3 as well as 64 in a robustness experiment (more details on the latter experiment will be provided in Section 3.3.1.3). Each experimental session took 1.5 hours.

All experimental treatments are based on the acyclical 11-20 game above, with an experimental currency where one token is worth 15 euro cents in Experiment 1 and 12.5 in Experiments 2 and 3. The exact sequences of treatments used in each session and experiment are provided in Appendix 3.7.2.2. Each subject in the experiments was anonymously paired with a new opponent after every iteration of the game. To focus on initial responses and to avoid learning from taking place, the subjects received no feedback after their play, and they only observed their earnings at the end of the session. As is standard in the literature on initial responses (see, e.g., Costa-Gomes et al. (2001) and Costa-Gomes and Crawford (2006)), subjects were paid randomly, and therefore did not have any mechanism for hedging against risk by changing their actions. Subjects were informed of the payment method before starting the experiment. Lastly, subjects received no information concerning other subjects' earnings. This serves to avoid that subjects focus on goals other than monetary incentives, such as

Goeree and Holt. (2004)) outperforms the baseline level-k model without noise.

defeating the opponent or winning for its own sake. The instructions of the experiment were given in Spanish; the English translation and the details on the pool of subjects, the earnings and the logistics of the experiments are in Appendix 3.7.2.

3.2.3 Subjects' Classifications

In Experiment 1, we divided the pool of subjects into two groups, according to two criteria (with 3 sessions of 20 subjects each) designed to be indicative of subjects' cognitive sophistication. The first criterion, referred to as the exogenous classification, separates subjects by their degree of study. Half of the students are drawn from humanities, and the other half from math and sciences. Subjects are then made aware of their own classification by being labeled as either 'humanities' or 'math and sciences'. In the endogenous classification, subjects are not separated by degree of study; instead, they are separated by a test that they take at the beginning of the experiment. This test consists of a centipede game, a pirates game and a simplified version of mastermind (see Appendix 3.7.2 for details). The top half of the subjects are labeled as high, and the bottom half as low. Here as well, the subjects are made aware of their own label, 'high' or 'low'. The details are provided in Appendix 3.7.2.

We use two different classifications for the following reason. In the exogenous classification, the labels are informative of a long-lasting, persistent and salient indicator of the subjects' cognitive sophistication.[6] The downside, however, is that it is a coarser notion of cognitive sophistication, and is not specifically linked to subjects' ability to reason strategically. In the endogenous classification instead, the labels are assigned based on a short test and may not necessarily have the persistent strength of the exogenous classification, but the

[6]This is the case especially when considering that at the university from which the subjects are drawn, there is a significant difference in the entry grades between those taking the fields grouped as humanities in our classification and those taking the fields grouped as math and science.

advantage is that the test specifically targets game theoretic reasoning, and so may induce sharper beliefs among the subjects concerning their relative sophistication compared to their opponents.

In Experiments 2 and 3, subjects were not separated by cognitive sophistication. They first took an expanded version of the test used in the endogenous classification described above. This version includes the muddy faces game in addition to the others (see Appendix 3.7.2). Then, only the subjects in the middle half of the distribution participated in the treatments below. The middle group was identified based on a pre-specified cutoff that was based on the test scores from Experiment 1. Subjects in this group were not given any information about their performance on the test before the treatments were administered. Those subjects who had high or low test results were given tasks to occupy them to prevent disturbances from subjects leaving during the experiment. This particular procedure was carried through for the following reasons. First, taking the test beforehand places the subjects in a similar condition to those of the endogenous classification treatments. Second, the treatments in Experiments 2 and 3 are designed to focus more on the middle part of the distribution, since in Experiment 1 the treatments of the exogenous classification focused on the tails of the distribution and those of the endogenous classification split subjects along the median. We do so to obtain additional information about a different segment of the distribution. We refer to the subjects in Experiments 2 and 3 as being in the *'unlabeled' classification*.

3.3 Experimental Design

3.3.1 Experiment 1: Treatments

We present next the treatments of Experiment 1. In Section 3.3.1.1 we review the baseline treatments in Alaoui and Penta (2016a), which test the basic premises of the EDR model by varying incentives or beliefs for both agents at the same time. These treatments, however,

Baseline Treatments	Opponent's label compared to own	Own payoffs	Opponent's payoffs	Replacement of opponent's opponent
Homogeneous [Hom]	same	Low	Low	No
Heterogeneous [Het]	different	Low	Low	No
Higher-Order Beliefs [HOB]	different	Low	Low	Yes
Homogeneous-high [Hom+]	same	High	High	No
Heterogeneous-high [Het+]	different	High	High	No
Higher-Order Beliefs-high [HOB+]	different	High	High	Yes

Table 3.1: Summary of the baseline treatments

do not disentangle whether subjects' change in behavior is due to changing their own incentives or to changing the incentives of the opponents, and whether subjects' choices are mainly driven by their beliefs about the opponents, or by their own cognitive limitations. The new treatments, designed to disentangle these effects, are introduced in Sections 3.3.1.2 and 3.3.1.3.

3.3.1.1 Baseline Treatments

AP's baseline treatments, summarized in Table 3.1, are designed to implement the two sets of comparative statics (on incentives and beliefs) we discussed above.

Varying Incentives. To vary subjects' incentives to reason, we consider two versions of the game: in the 'low payoffs' treatment, we set $x = 20$; in the 'high payoffs' treatments, we let $x = 80$. Note that this change does not affect the level-k actions, irrespective of whether the level-0 is specified as 20 or as the uniform distribution. It only increases the rewards for players who stop at the 'correct' round of reasoning, and hence the 'incentives to reason'.[7]

Varying Beliefs. To vary agents' beliefs, for both specifications of payoffs and for both the classification criteria discussed in Section 3.2.3, subjects in each treatment are given information concerning

[7] Alaoui and Penta (2018) provide axiomatic foundations to this assumption of the EDR model.

their opponent's label. They play the baseline game against someone from their own label (*homogeneous treatment*) and against someone from the other label (*heterogeneous treatment*). For instance, for the exogenous classification, a student from math and sciences (resp., humanities), is told in the homogeneous treatment that his opponent is a student from math and sciences (humanities). In the heterogeneous treatment, he is told that the opponent is a student from humanities (math and sciences). Identical instructions are used for the endogenous classification, but with 'high' and 'low' instead of 'math and sciences' and 'humanities', respectively.

The homogeneous and heterogeneous treatments are designed to test whether the behavior of the subjects varies with the sophistication of the opponent. The next treatment is designed to test whether the subjects believe that the behavior of their opponents also changes when they face opponents of different levels of sophistication. To do so, we consider a *higher order beliefs treatment*: A 'math and sciences' subject, for instance, is given the following instructions: "[...] two students from humanities play against each other. You play against the number that one of them has picked."

In the following, we let [Hom], [Het] and [HOB] denote, respectively, the homogeneous, heterogeneous and higher-order beliefs treatments when payoffs are low, and [Hom+], [Het+] and [HOB+] the corresponding treatments when payoffs are high.

3.3.1.2 Relaxing Cognitive Bounds: the post-Tutorial Treatments

We explain next the new treatments designed to identify whether or not subjects play according to their own (binding) cognitive bound in treatments [Hom] and [Het].

Consistent with the intuitive idea of the *tutorial method* discussed in the introduction, after having administered the baseline treatments of Section 3.3.1.1, we exposed all eighty subjects from four of the six sessions (two for the endogenous and two for the exogenous classifica-

The Tutorial Treatments	Opponent's label compared to own	Own payoffs	Opponent's payoffs	Own Tutorial	Opponent's Tutorial
Tutorial [Tut]	–	Low	Low	Yes	Yes
Asymm. Tutorial-Homog. [AT-Hom]	same	Low	Low	Yes	No
Asymm. Tutorial-Heterog. [AT-Het]	different	Low	Low	Yes	No

Table 3.2: Summary of the post-tutorial treatments.

tions) to a 'game theory tutorial'. This tutorial explains how, through the chain of best replies, 'infinitely sophisticated and rational players' would play $(11, 11)$:

> According to game theory, if the players are infinitely rational, then the game should be played in the following way. Both players should say 11.
>
> *Explanation*: Suppose the two players are named Ana and Beatriz. If Ana thinks Beatriz plays 20, then Ana would play 19. But then Beatriz knows that Ana would play 19, so she would play 18. Ana realizes this, and so she would play 17.... they both follow this reasoning until both would play 11. Notice that if Beatriz says 11, then the best thing for Ana is to also say 11.

We then proceed with three new (post-tutorial) treatments, each repeated twice, and summarized in Table 3.2.

In treatment [Tut], we instruct each subject to play the baseline game (with low payoffs) against another subject who has also been given the same tutorial, with no information about his label. In the 'asymmetric tutorial-homogeneous' treatment [AT-Hom], we instruct the subjects who had previously received the tutorial to play the baseline game against a player of the same label who had *not* received the tutorial (that is, as in the baseline homogeneous treatment [Hom]). Analogously, the 'asymmetric tutorial-heterogeneous' treatment [AT-Het] contains the same instructions but with the subjects facing an opponent from a different label (as in baseline treatment [Het]).

Hence, subjects essentially face the same opponents in treatments [AT-Hom] and [Hom] (and in [AT-Het] and [Het]), but the tutorial ensures that their cognitive bounds are not binding in treatments [AT-Hom] and [AT-Het]. Hence, if their cognitive bound was *not* binding in (pre-tutorial) treatments [Hom] and [Het], then the distributions of actions in (post-tutorial) treatments [AT-Hom] and [AT-Het] should be the same as in [Hom] and [Het], respectively.

3.3.1.3 Reasoning about Others' Incentives: Asymmetric Payoffs Treatments

In this section we explain the treatments designed to disentangle the effects of increasing payoffs on subjects' own cognitive bound from their reasoning about others' incentives to reason. As discussed in the introduction, this question is inherently related to the cognition-belief dichotomy (the object of the treatments in Section 3.3.1.2), but it is more directly motivated by the basic premise of the EDR.

In the design of treatments [Hom+], [Het+] and [HOB+], relative to [Hom], [Het], [HOB], we increase the payoff for undercutting the opponent for both players in the game. Thus, the shifts in the distributions towards lower numbers observed in Alaoui and Penta (2016a) (Section 4) may conflate two distinct effects. The first effect is the possible increase in the cognitive bound of player i, and the second is the change in i's beliefs about j's cognitive bound due to the change in j's incentives. Both effects would determine an increase in the behavioral level, hence a shift of the distribution towards lower actions. The following treatments, summarized in Table 3.3, are aimed at disentangling the two effects, and testing whether subjects in our experiment reason about their opponents' incentives independently of their own.

As discussed earlier, the intuitive idea is to increase the stakes for one player without changing the other player's. To address the problem of higher-order beliefs discussed in the introduction, however, in these treatments we apply the *replacement method* to the game with

Asymmetric Payoffs Treatments	Opponent's label compared to own	Own payoffs	Opponent's payoffs	Replacement of opponent's opponent
Asymm. Payoffs-Homogeneous [AP-Hom]	same	High	Low	Only payoffs
Asymm. Payoffs-Heterogeneous [AP-Het]	different	High	Low	Both label and payoffs

Table 3.3: Summary of the asymmetric payoff treatments.

asymmetric payoffs. That is, in treatments [AP-Hom] and [AP-Het] agents play the high-payoff game against the number chosen by an opponent from their own or the other label respectively in the low payoffs treatment [Hom]. Hence, the exercise is of a similar spirit to treatment [HOB], in which subjects play against the number chosen by an opponent from a different label who is engaged in treatment [Hom]. Both treatments are administered after the baseline treatments to all forty subjects from two sessions, one exogenous and one endogenous, and each is repeated three times.

Note that these treatments add a further layer of complexity, since the individual is told in treatment [AP-Hom] (resp., [AP-Het]) that he is playing the high-payoff game against the number chosen by an opponent of the same (other) label himself playing the low payoff game against an opponent of their own label. Treatment [AP-Het] is especially complex: for player i, both the payoffs and the label of i's opponent *and* of the opponent's opponent are different from i's own payoff and label.

By comparing treatments [AP-Hom] and [AP-Het] with treatments [Hom] and [HOB] and with treatments [Hom+] and [HOB+], we can disentangle the two effects mentioned above. The shift from [Hom] to [AP-Hom] (and from [HOB] to [AP-Het]), due solely to the increase of each subject's own payoffs and not his opponent's, may be attributed to the increase of subjects' own cognitive bound. It should be observed only if the cognitive bound in treatments [Hom] and [HOB] had been binding; the further shift from [AP-Hom] to [Hom+] (and from [AP-Het] to [HOB+]) instead can be imputed to the increase in subjects'

beliefs about their opponents' behavior due to the increase of their payoffs.

Robustness experiment. When comparing [Hom] to [AP-Hom], one potential issue is that the replacement method could remove social preferences, if those were present. Specifically, social preferences might impact [Hom] but not [AP-Hom] so that when we compare the behavior across these treatments, we might confound differences in levels with social preference differences. To assess the robustness of our findings, we designed an extra experiment. We conducted a robustness experiment that also contains a new homogeneous treatment with replacement, [Hom-Rep], in order to make [Hom] and [AP-Hom] more comparable.[8] In [Hom-Rep], the subject played against another subject from their own group who in turn was playing yet another player from their own group. This means that there now is replacement both for the [Hom-Rep] benchmark and the replacement treatment [AP-Hom]. For comparability, the experiment also contained the basic [Hom] treatment. We conducted the experiment with 64 subjects at Universitat Pompeu Fabra, in May 2019.

3.3.2 Experiment 2: Unlabeled Variations

A possible concern with the design of Experiment 1 is that the joint presence of beliefs and payoff treatments may increase the complexity of the experimental instructions, adding noise to the results. Experiment 2 is designed to simplify the cognitive load of the instructions by replicating the main treatments without label distinctions. For this reason, subjects in this experiment are not given any information on their own or the opponents' performance in the cognitive sophistication test. This change simplifies the instructions, particularly in the post-tutorial treatments and asymmetric payoff treatments. This is because a subject need not keep track at the same time of whether the opponent has taken the tutorial (for the former case, or has different

[8]We thank an anonymous referee for suggesting this treatment.

Baseline Treatments	Own payoffs	Opponent's payoffs	Own Tutorial	Opponent's tutorial	Replacement of opponent's opponent
Unlabeled [Un]	Low	Low	No	No	No
Unlabeled-high [Un+]	High	High	No	No	No
Tutorial-Unlabeled [Tut-Un]	Low	Low	Yes	Yes	No
Asymm.Tut.-Unlab. [AT-Un]	Low	Low	Yes	No	Tutorial only
Asymm.Payoffs-Unlab. [AP-Un]	High	Low	No	No	Payoffs only

Table 3.4: Summary of the treatments in Experiment 2.

payoffs for the latter) and of his possibly different label as well as the higher order beliefs over the opponent's opponent on both dimensions.

All 60 subjects participate in all treatments. As usual, they do not receive feedback, and they are paid at random based on their behavior on a subset of the treatments (see Appendix A for the exact sequence of treatments, payment details and wording of instructions).

Treatments. The baseline unlabeled treatment, denoted [Un], consists of the subjects playing the baseline acyclical 11-20 game with extra reward $x = 20$ (see Section 3.2.1), and Unlabeled-high [Un+] consists of the subjects playing the high payoffs game with extra reward $x = 80$.

The tutorial-unlabeled treatment [Tut-Un] is identical to post-tutorial treatment [Tut] except that subjects are not given any label. In Treatment [AT-Un], subjects who have seen the tutorial play against the action chosen by an opponent in pre-tutorial treatment [Un]. This treatment serves to fix both the subject's beliefs that his opponent has not seen the tutorial and his beliefs that his opponent's opponent has not seen the tutorial either (and so on for all higher-order beliefs). Lastly, we have adapted the asymmetric payoffs treatments in the same way, so that Asymmetric Payoffs-Unlabeled [AP-Un] is identical to [AP-Hom] and [AP-Het] without label information. These treatments are summarized in Table 3.4.

3.3.3 Experiment 3: Unlabeled Variations with preliminary semi-Tutorial

One possible concern with the previous experiments could be that the tutorial induces level-k thinking. If the subjects were not adopting that kind of reasoning in the pre-tutorial treatments, the difference between the pre- and post-tutorial behavior could in part be due to this form of priming. If that were the case, interpreting the changes in behavior as being due to relaxing a possible cognitive constraint of the subjects may be problematic, as it could be conflating different factors.[9]

To address these concerns, Experiment 3 replicates Experiment 2, but precedes the administration of the treatments with instructions designed to keep the priming as similar as possible between the tutorial and non-tutorial treatments. Specifically, we modified the instructions before the first game to have the following explanation, and kept all else identical:

> If you think that your opponent will choose 20, or any number with equal likelihood, then the action that will maximize your earnings would be to choose 19. If you think that following this observation means that your opponent chooses 19, then the action that maximizes your earnings would be to choose 18.

Arguably, these instructions prime level-k reasoning in equal measure as the tutorial treatment, in that they make explicit the start of the chain of reasoning. They are effectively identical to the tutorial except that they do not spell out the *entire* path of reasoning all the way to 11. Hence, if such priming were the main driving force in the results for Experiments 1 and 2, then it would not be the case here. Since, in Experiment 3, all subjects were already primed to think according to level-k when playing the first treatment [E3-Un], changes observed

[9]We are thankful to an anonymous referee for this observation.

Baseline Treatments	Own payoffs	Opponent's payoffs	Own Tutorial	Opponent's tutorial	Replacement of opponent's opponent
Unlabeled [E3-Un]	Low	Low	No	No	No
Unlabeled-high [E3-Un+]	High	High	No	No	No
Tutorial-Unlabeled [E3-Tut-Un]	Low	Low	Yes	Yes	No
Asymm.Tut.-Unlab. [E3-AT-Un]	Low	Low	Yes	No	Tutorial only
Asymm.Payoffs-Unlab. [E3-AP-Un]	High	Low	No	No	Payoffs only

Table 3.5: Summary of the treatments in Experiment 3.

between the pre- ([E3-Un]) and post-tutorial responses ([E3-AT-Un]) should not stem from priming but from changes in their cognitive levels.

In addition, subjects were asked to complete a Theory of Mind test (Stiller and Dunbar (2007)) at the end of the experiment to provide a measure of their ability to place themselves in the mind of another person and to thus form higher order beliefs (we further explain this test in Section 3.5.3).

Treatments. The treatments and logistics are just the same as in Experiment 2, with the only difference being that all 34 subjects were exposed to the semi-tutorial at the very beginning of the session. We thus maintain the corresponding labels, preceded by "E3" to denote they refer to the variation of Experiment 3. These treatments are summarized in Table 3.5.

Table 3.6 summarizes all treatments implemented across all three experiments.

3.4 The EDR model

The basic idea of the EDR model is that a subject's 'level of play', or *behavioral level*, may be endogenous due to two related mechanisms. First, given a subject's understanding of a game (his *cognitive bound*,

Payments	Same label (Homogeneous)		Different labels (Heterogeneous)		No label (Unlabeled)			
	No tutorial	Own tutorial	No tutorial	Own tutorial	No tutorial	Own tutorial	Both tutorial	Semi-tutorial
Both low	[Hom]	[AT-Hom]^{tr}	[Het] [HOB]^{lr}	[AT-Het]^{tr,lr}	[Un]	[AT-Un]^{tr} [E3-AT-Un]^{tr}	[Tut] [Tut-Un] [E3-Tut-Un]	[E3-Un] [E3-AT-Un]^{tr} [E3-Tut-Un]
Both high	[Hom+]		[Het+] [HOB+]^{lr}		[Un+]			[E3-Un+]
Own high, opponent's low	[AP-Hom]^{pr}		[AP-Het]^{pr,lr}		[AP-Un]^{pr}			[E3-AP-Un]^{pr}

Superscripts next to the treatments indicate replacement of opponent's opponent's payoff ([.]^{pr}), label ([.]^{lr}) or tutorial ([.]^{tr}).

Table 3.6: Summary of all treatments over all experiments.

or *capacity*), his 'behavioral level' may vary with his beliefs about the opponent: e.g, even if a subject understands up to five iterations of the level-k reasoning, he may sometimes play as a level-5 (e.g., choose 15 in the 11-20 game), but sometimes play as a level-3, if he thinks the opponent would play as a level-2. But clearly, it is a matter of definition that one never plays according to a higher level than one's own capacity. Hence, if $\hat{k}_i$ is the cognitive bound of subject i, his possible 'behavioral levels' are $k_i \leq \hat{k}_i$. And for the same reason, i's perception of the opponent's capacity, $\hat{k}_j^i$, is also bounded by his own: $\hat{k}_j^i < \hat{k}_i$.[10]

The second dimension of endogeneity is that the understanding of a game (i.e., the capacity), may itself vary with a player's stakes in the game. For instance, in the acyclical 11-20 game, in which $x_i \geq 20$ denotes the extra reward that player i gets for being exactly one below the opponent, the EDR model implies that agent i's capacity $\hat{k}_i$ (as well as his perception of the opponent's capacity, $\hat{k}_j^i$) is weakly

[10] A different modeling choice would be to assume that the players first consider the sophistication of their opponent, and stop reasoning as soon as they believe they have exceeded it if the opponent is less sophisticated; that is, as soon as player i reaches step $\hat{k}_j^i + 1$. This would lead to a different interpretation in that own capacity and beliefs would coincide, but it would be behaviorally equivalent (cf. Alaoui and Penta (2016a)).

increasing in x_i, and that $\hat{k}_j^i$ is weakly increasing in the opponent's extra reward, x_j. We introduce next a simple version of the EDR model which formalizes these ideas, as well as the interactions between individuals' beliefs and incentives to reason, and derive its predictions for the treatments in the experiments above. All proofs of our results are in Appendix 3.7.1.

3.4.1 Baseline Model

Own understanding, Costs and Values: The endogeneity of players' capacities is modeled as stemming from a cost-benefit analysis: costs represent players' cognitive abilities; the benefits instead only depend on the game's payoffs, such as the x parameter in the 11-20 game. Formally, fixing the game payoffs, let $v_i : \mathbb{N} \to \mathbb{R}_+$ denote the value of reasoning, where $v_i(k)$ represents i's value of doing the k-th round of reasoning, given the previous $k-1$ rounds. The cognitive ability of agent i is represented by a cost function $c_i : \mathbb{N} \to \mathbb{R}_+$, where $c_i(0) = 0$ and $c_i(k)$ denotes i's incremental cost of performing the k-th round of reasoning. We say that cost function c' is 'more (resp. less) sophisticated' than c'', if $c'(k) \leq c''(k)$ (resp., if $c'(k) \geq c''(k)$) for every k. For any $c_i \in \mathbb{R}_+^{\mathbb{N}}$, we denote by $C^+(c_i)$ and $C^-(c_i)$ the sets of cost functions that are respectively 'more' and 'less' sophisticated than c_i.

For the 11-20 games of our experiments, the general assumptions of the EDR model imply that: (i) the value of reasoning only depends on i's payoffs of the game; (ii) for every i and k, the value $v_i(k)$ is (weakly) increasing in x_i and constant in x_j; (iii) $v_i = v_j$ if $x_i = x_j$; (iv) the costs of reasoning are constant throughout all variations of the game in all non post-tutorial treatments.[11] To obtain testable predictions in our experiments, further assumptions on subjects' beliefs and the

[11] These assumptions are all implied by the general, "detail free", EDR model (Alaoui and Penta (2016a)), as well as consistent with the axiomatic foundations provided in Alaoui and Penta (2018). We refer to those papers for discussions of the conceptual significance of these assumptions.

tutorial's effects are needed, and will be discussed below.

To allow for the case, as in our asymmetric payoff treatments, that j's opponent is not i but a low-payoff version of it, besides values v_i and v_j we also introduce $v_{i(j)}$, to denote the value of reasoning of j's opponent. This value may coincide with i's own value ($v_{i(j)} = v_i$) – as in the standard treatments – or not – as in the asymmetric payoffs treatments ($v_{i(j)} \neq v_i$). More specifically, in general we assume that (v) $v_{i(j)}$ is equal to what would be i's value if his extra reward x_i were equal to that of j's opponent, $x_{i(j)}$.

For later reference, we define a mapping $\mathcal{K} : \mathbb{R}_+^{\mathbb{N}} \times \mathbb{R}_+^{\mathbb{N}} \to \mathbb{N}$ such that, $\forall\, (c, v) \in \mathbb{R}_+^{\mathbb{N}} \times \mathbb{R}_+^{\mathbb{N}}$,

$$\mathcal{K}(c, v) := \min \{k \in \mathbb{N} : c(k) \leq v(k) \text{ and } c(k+1) > v(k+1)\}, \tag{3.1}$$

where $\mathcal{K}(c, v) = \infty$ if the set in equation (3.1) is empty. In words, this mapping identifies the first intersection between the value v and the cost c.

Player i's *cognitive bound* is the value that this function takes at (c_i, v_i):

$$\hat{k}_i = \mathcal{K}(c_i, v_i). \tag{3.2}$$

Beliefs and Others' Understanding: To distinguish players' cognitive and behavioral levels, the EDR model also specifies beliefs about the opponent's costs, as well as higher order beliefs, which are then used to derive i's beliefs about the opponent's cognitive bound, his beliefs about j's beliefs about i's bound, and so on. In the general EDR model, such beliefs are modeled through *cognitive type spaces*, which can be used to represent arbitrary belief hierarchies over players' costs (cf. Alaoui and Penta, 2016a). Here, however, it suffices to focus on the simpler case of *second-order types with degenerate beliefs*. These beliefs are pinned down by (1) i's cost function, c_i, (2) i's beliefs about j's cost function, c_j^i, and (3) i's beliefs about j's beliefs over i's cost function, c_i^{ij} (which may or may not be such that $c_i^{ij} = c_i$ – the further special case of "common belief "types).[12] A *type* in the

[12]We refer to Alaoui and Penta (2016a) for the general case with non-degenerate

following is thus a triple $t_i = (c_i, c_j^i, c_i^{ij})$. With this notation, we define i's beliefs about j's cognitive bound (given his own bound $\hat{k}_i$, and his beliefs about j's cost, c_j^i) as:

$$\hat{k}_j^i = \min\left\{\hat{k}_i - 1, \mathcal{K}(c_j^i, v_j)\right\}. \tag{3.3}$$

The minimum operator represents the idea that i's beliefs over j's capacity are bounded by his own cognitive bound, $\hat{k}_i$, which effectively only uncovers the understanding of levels $k_i < \hat{k}_i$.

Player i's beliefs over $\hat{k}_j^i$, j's cognitive bound, do not depend on c_i^{ij}, but his beliefs over k_j^i, j's level of play, do. That is, k_j^i can be below $\hat{k}_j^i$ if i believes that his sophistication is underestimated by j. Put differently, player i attempts to place himself in the mind of j, and views j's beliefs over his own cognitive bound, $\hat{k}_i^{ij}$, to be:

$$\hat{k}_i^{ij} = \min\left\{\hat{k}_j^i - 1, \mathcal{K}(c_i^{ij}, v_{i(j)})\right\}. \tag{3.4}$$

The minimum operator in (3.4) reflects the fact that, just as i's beliefs over j's capacity are bounded by $\hat{k}_i$, so his beliefs about j's beliefs over i's capacity are bounded by what i thinks j's capacity is. The fact that $v_{i(j)}$ is used reflects the idea that i understands that j's behavior is based on his reasoning about his opponent, which – depending on the treatment – may be i himself ($v_{i(j)} = v_i$) or a replaced version of it with different payoffs ($v_{i(j)} \neq v_i$).

Behavior: For player i's beliefs over j's behavior, player i expects j to play according to level $\hat{k}_i^{ij} + 1$, provided that i thinks that j is capable of conceiving of such a level, which is the case if $\hat{k}_i^{ij} + 1 \leq \hat{k}_i - 1$. Otherwise, i thinks that j is limited by his own cognitive bound. Hence, for a general second-order type, i's perception of j's behavioral bound is:

$$k_j^i = \min\left\{\hat{k}_i^{ij} + 1, \hat{k}_i - 1\right\}. \tag{3.5}$$

beliefs, and an explanation of how second-order types map to the language of cognitive type spaces of the general model.

Player i then best responds to k_i^j, and hence his *behavioral level* is $k_i = k_j^i + 1$. In the acyclical 11-20 game, the associated actions are $a_i^{k_i} = 20 - k_i$ if $k_i \leq 9$, and $a_i^{k_i} = 11$ otherwise.

Letting x_i, x_j, and $x_{i(j)}$ denote, respectively, the extra reward in the 11-20 game received by player i, j, and j's opponent, in the following proposition we refer to the 11-20 game with $x_i = x_j = x_{i(j)} = 20$ as the *low payoff game,* to the case with $x_i = 80 \neq x_j = x_{i(j)} = 20$ as the *asymmetric payoff game,* and to the case $x_i = x_j = x_{i(j)} = 80$ as the *high payoff game.*

Proposition 1. *Fix a second-order type $t_i = (c_i, c_j^i, c_i^{ij})$. Then:*

1. *If $c_j^i, c_i^{ij} \in C^+(c_i)$, then the cognitive bound is binding in the low-payoff game, and $k_i = \hat{k}_i$. In the asymmetric payoff game, both k_i and $\hat{k}_i$ are (weakly) higher than in the low payoff game; the cognitive bound may or may not be binding anymore, and $k_i \leq \hat{k}_i$ may also hold with strict inequality. In the high payoff game, $\hat{k}_i$ remains the same as in the asymmetric payoff game, and it is such that $k_i = \hat{k}_i$; k_i may increase or stay the same, but it increases* only if *the cognitive bound* was not *binding in the asymmetric payoff game.*

2. *If $c_j^i \in C^-(c_i)$, or if $c_j^i \in C^+(c_i)$ and $c_i^{ij} \in C^-(c_i)$, then the cognitive bound may or not be binding in the low-payoff game, and $k_i \leq \hat{k}_i$ may also hold with strict inequality. In the asymmetric payoff game, $\hat{k}_i$ is (weakly) higher than in the low payoff game; k_i may increase or stay the same, but it increases* only if *the cognitive bound* was *binding in the low-payoff game. In the high-payoff game, $\hat{k}_i$ is the same as in the asymmetric payoff game, and may or may not be binding; k_i may increase or stay the same, but it increases* only if *the cognitive bound* was not *binding in the asymmetric payoff game.*

3. *For any (c_j^i, c_i^{ij}) and $(x_i, x_j, x_{i(j)})$, replacing c_i with some lower-cost $c_i' \in C^+(c_i)$ always induces a (weakly) higher $\hat{k}_i$; k_i may*

increase or stay the same, but it increases only if *the cognitive bound* was *binding in the first place.*

3.4.2 Identification Assumptions and Predictions for the Experiments

As briefly mentioned, to obtain testable predictions in our experiments we need to append the EDR's model assumptions (i)-(v) on the properties of the costs and value of reasoning, with identification assumptions on how the treatment variations impact players' beliefs and costs of reasoning (which are the exogenous *types* in the EDR model).

The first identification assumption that we introduce formalizes the effects of the tutorial as effectively eliminating the costs of performing any step of level-k reasoning:[13]

IA.1: in the post-tutorial treatments, $c_i(k) = 0$ for all k.

The next assumption restricts the way that subjects' beliefs vary from one treatment to the other. While alternatives are possible in practice, in order to ensure that the model has bite we impose the most restrictive assumptions on beliefs which is sensible in the present context:

IA.2: For all treatments other than [Tut], [Tut-Un] and [E3-Tut-Un], subject i's first-order beliefs, c_j^i, only depend on the label of the opponent, and his second order beliefs c_i^{ij} only depend on the label of the opponent's opponent.

For the unlabeled treatments of Experiments 2 and 3, IA.2 effectively implies that beliefs c_j^i and c_i^{ij} are constant throughout the experiment, except for treatments [Tut-Un] and [E3-Tut-Un]. The

[13]While it may well be that the cost is not shifted all the way down to 0 in the post-tutorial treatments, this assumption ties our hands maximally, and allows for clean predictions.

Equivalence classes for c_i in the treatments of Exp.2 & Exp.3	Equivalence classes for c_j^i in the treatments of Exp.2 & Exp.3	Equivalence classes for c_i^{ij} in the treatments of Exp.2 & Exp.3
[Un]=[Un+]=[AP-Un] [Tut-Un]=[AT-Un]	[Un]=[Un+]=[AP-Un]=[AT-Un] [Tut-Un]	[Un]=[Un+]=[AP-Un]=[AT-Un] [Tut-Un]
[E3-Un]=[E3-Un+]=[E3-AP-Un]	[E3-Un]=[E3-Un+]= =[E3-AP-Un]=[E3-AT-Un]	[E3-Un]=[E3-Un+]= =[E3-AP-Un]=[E3-AT-Un]
[E3-Tut-Un]=[E3-AT-Un]	[E3-Tut-Un]	[E3-Tut-Un]

Table 3.7: For any subject $t_i = (c_i, c_j^i, c_i^{ij})$, the table shows the classes of treatments that generate the same c_i, c_j^i or c_i^{ij} in Experiments 2 and 3

Equivalence classes for c_i in the treatments of Exp.1	Equivalence classes for c_j^i in the treatments of Exp.1	Equivalence classes for c_i^{ij} in the treatments of Exp.1
[Hom]=[Hom+]=[Het]= =[Het+]=[HOB]=[HOB+]= =[AP-Hom]=[AP-Het]	[Hom]=[Hom+]=[AT-Hom]	[Hom]=[Hom+]= =[Het]=[Het+]=[AP-Hom]
[Tut]=[AT-Hom]=[AT-Het]	[Het]=[Het+]=[HOB]= =[HOB+]=[AP-Het]=[AT-Het] [Tut]	[HOB]=[HOB+]= =[AP-Het]=[AT-Het] [Tut]

Table 3.8: For any subject $t_i = (c_i, c_j^i, c_i^{ij})$, the table shows the classes of treatments that generate the same c_i, c_j^i or c_i^{ij} in Experiment 1

reason why the latter treatments are treated differently is that in such treatments subjects are informed that the opponent also took the tutorial, and hence – consistent with the spirit of IA.1 – we assume that beliefs $c_j^i(k)$ and $c_i^{ij}(k)$ also get lower for every k. As previously discussed, however, our main interest lies in comparing the pre- and post-tutorial treatments, not in the [Tut] treatment per se. For those pre- and post-tutorial comparisons, in which opponents are effectively the same, IA.2 implies that c_j^i and c_i^{ij} remain unchanged. Table 3.7 summarizes the implications of identification assumptions IA.1 (first column) and IA.2 (second and third columns) for the treatments in Experiments 2 and 3.

The implications of IA.2 for the labeled treatments in Experiment

1 are more complicated. They imply, for instance, that c_j^i remains the same in the [Het] and [HOB] treatments, but may change moving from [Hom] to [Het]; in contrast, c_i^{ij} is the same in the [Hom] and [Het] treatments, but may change moving from [Het] to [HOB] (and similarly for the high payoff versions of these treatments). Finally, both c_j^i and c_i^{ij} remain the same moving from treatment [X] to [X+], for all X $\in$ {Hom, Het, HOB}. The overall implications of assumptions IA.1-2 for the treatments of Experiment 1 are summarized in Table 3.8.

Under these two basic identification assumptions, the EDR model implies clear predictions on the comparisons for most of our treatments, which we summarize in Proposition 2. In the following, we let F_X^l denote the cumulative distribution of actions $a \in \{11, ..., 20\}$ in treatment $[X]$ for label $l \in \{I, II, *\}$ (where "$l = *$" means *unlabeled*, as in Experiments 2 and 3), and denote by $\succsim$ (resp., $\succ$) the weak (resp., strict) first order stochastic dominance relation.[14]

Proposition 2. *For any distribution over subjects' types $t_i = (c_i, c_j^i, c_i^{ij})$ which satisfy identification assumptions IA.1-2, under the maintained assumptions of the EDR model (Section 3.4.1), the following holds:*

1. *In Experiment 1, for each $l \in \{I, II\}$: $F_X^l \succsim F_{X+}^l$ for all $X \in \{Hom, Het, HOB\}$.*

2. *In Experiment 1, for each $l \in \{I, II\}$: (i) $F_{Hom}^l \succsim F_{AP-Hom}^l \succsim F_{Hom+}^l$, with $F_{Hom} \succ F_{AP-Hom}^l$ only if $\hat{k}_i$ was binding in [Hom] for some i; and (ii) $F_{HOB}^l \succsim F_{AP-Het} \succsim F_{HOB+}^l$, with $F_{HOB}^l \succ F_{AP-Het}$ only if $\hat{k}_i$ was binding in [HOB] for some i.*

3. *In Experiment 1, for each $l \in \{I, II\}$: $F_{Hom}^l \succsim F_{AT-Hom}^l$, and $F_{Het}^l \succsim F_{AT-Het}^l$, each strictly only if $\hat{k}_i$ was binding for some i in [Hom] and [Het], respectively.*

[14] Given two cumulative distributions $F(x)$ and $G(x)$, we say that F (weakly) first order stochastically dominates G, written $F \succsim G$, if $F(x) \leq G(x)$ for every x. $F \succ G$ if $F \succsim G$ and $F(x) < G(x)$ for some x

4. *In Experiments 2 and 3: (i)* $F^*_{Un} \succsim F^*_{AP-Un} \succsim F^*_{Un+}$*, with* $F^*_{Un} \succ F^*_{AP-Un}$ *only if* $\hat{k}_i$ *was binding in [Un] for some* i*; and (ii)* $F^*_{E3-Un} \succsim F^*_{E3-AP-Un} \succsim F^*_{E3-Un+}$*, with* $F^*_{Un} \succ F^*_{E3-AP-Un}$ *only if* $\hat{k}_i$ *was binding in [E3-Un] for some* i*.*

5. *In Experiments 2 and 3:* $F^*_{Un} \succsim F^*_{AT-Un}$ *and* $F^*_{E3-Un} \succsim F^*_{E3-AT-Un}$*, each strictly only if* $\hat{k}_i$ *was binding for some* i *in [Un] and [E3-Un], respectively.*

The remaining identification assumptions only concern the treatments in Experiment 1.

IA.3: For the labeled treatments of Experiment 1, we assume that individuals commonly believe that label I players are more sophisticated than label II. Formally, for label I individuals: if $l_i = I$, $c_j^{i,[Het]} \in C^-(c_j^{i,[Hom]})$, $c_i^{ij,[HOB]} \in C^-(c_i^{ij,[Het]})$; For label II individuals: if $l_i = II$, $c_j^{i,[Hom]} \in C^-(c_j^{i,[Het]})$, $c_i^{ij,[Het]} \in C^-(c_i^{ij,[HOB]})$.

IA.4: For the labeled treatments of Experiment 1, we assume that label II individuals (i) always regard label I's as more sophisticated than they are, and (ii) they expect label I not to underestimate their sophistication. Formally: (i) $c_j^{i,[Het]} \in C^+(c_i)$ and (ii) $c_i^{ij,[Het]} \in C^+(c_i)$ whenever $l_i = II$.

Proposition 3. *For any distribution over subjects' types* $t_i = (c_i, c_j^i, c_i^{ij})$ *in Experiment 1 that satisfy assumptions IA.2-4, under the maintained assumptions of the EDR model (Section 3.4.1), the following holds:*[15]

1. *(i)* $F^I_{HOB} \succsim F^I_{Het} \succsim F^I_{Hom}$*; (ii)* $F^{II}_{Hom} \succsim F^{II}_{Het} \approx F^{II}_{HOB}$*; (iii)* $F^I_{HOB+} \succsim F^I_{Het+} \succsim F^I_{Hom+}$*; and (iv)* $F^{II}_{Hom+} \succsim F^{II}_{Het+} \approx F^{II}_{HOB+}$*.*

2. $F^I_{AT-Het} \succsim F^I_{AT-Hom}$ *and* $F^{II}_{AT-Hom} \succsim F^{II}_{AT-Het}$*.*

3. $F^I_{AP-Het} \succsim F^I_{AP-Hom}$ *and* $F^{II}_{AP-Hom} \succsim F^{II}_{AP-Het}$*.*

[15] If part (ii) of IA.4 were dropped, the only change to this Proposition is that parts the $\approx$ relation in parts (ii) and (iv) of point 1 would be weakened to $\succsim$.

4. *For label I, the increase in k_i from [HOB] to [AP-Het] should be at most one.*

Intuitively, to understand the effects of changing beliefs in the EDR model, when incentives are symmetric ($x_i = x_j$, as in the baseline treatments in Section 3.3.1.1), an individual's cognitive bound is binding if he regards his opponent as 'more sophisticated' (i.e., lower cost-of reasoning). Hence, when the incentives to reason are symmetric, individuals with higher costs of reasoning have a lower cognitive bound, which therefore is binding when playing against someone they regard as more sophisticated.

The reason for the asymmetric effect of higher-order beliefs in point 1 is that, for label II subjects, if their cognitive bound is binding in treatment [Het] (respectively, [Het+]) – in which they play against someone they regard as more sophisticated – then it would also be binding in treatment [HOB] (resp., [HOB+]) – in which their opponent may play according to an even deeper behavioral level – and therefore behavior should be the same in these treatments. Label I subjects, instead, would understand that label II subjects play according to a higher behavioral level in the [Het] than in the [Hom] treatment, and hence their behavioral level in treatment [HOB] may be lower than in treatment [Het], which in turn is lower than in [Hom]. *Hence, higher-order beliefs effects (i.e., comparing treatments [Het] and [HOB]) are possible, but they are one-sided: they should be observed, if at all, only for label I subjects.* These are precisely some of AP's main findings, which we summarize in Section 3.5.1.1.

The reason for point 4 is that, under the assumption that label I subjects regard label II subjects as having a higher cost of reasoning (less sophisticated) than themselves, in the [HOB] treatment the label II cognitive bound can be at most the same as label I's. If it is strictly less, then increasing label I's incentives should not affect their behavior, because their cognitive bound was binding in the first place. If instead the cognitive bounds were the same, then label I's behavior would change, but since the opponents' bound is the same in the two treatments, label I's behavioral level would only increase by one level.

Based on the logic above, and by IA.3-4 label I subjects are commonly regarded as 'more sophisticated' than label II, *we expect more label II subjects with binding cognitive bounds in treatment [Het] than in treatment [Hom] (and in [Het+] than in [Hom+], whereas the opposite would be true for label I subjects.* This implies the following proposition:

Proposition 4. *For any distribution over subjects' types $t_i = (c_i, c_j^i, c_i^{ij})$ in Experiment 1 that satisfy assumptions IA.1-4, under the maintained assumptions of the EDR model (Section 3.4.1), the following holds: For label II subjects (resp., label I) shift in behavior from [Het] to [AT-Het] is (weakly) larger (resp., smaller) than from [Hom] to [AT-Hom].*

Discussion of the Identification Assumptions: We briefly discuss here the possible weaknesses of the approach we follow, our reasoning behind our identification assumptions, and possible alternatives.

First of all, we note that identification assumptions are by their nature typically untestable within the same dataset, and our case is no exception. We have made these particular assumptions because we believe they are both natural and restrictive. For instance, assumption **IA.1** states that the tutorial reduces the costs to 0. If we allowed for the costs to be reduced by a smaller amount, then our predictions would be less sharp and the model would be less falsifiable. Assumption **IA.2**, which states that the subjects' beliefs in the pre-tutorial treatments depend only on the labels, could be replaced or relaxed. Our rationale for using it is that the labels are the only information that the subjects have, it is not ad-hoc, and it is restrictive enough to allow for relatively sharp predictions.

In the case of **IA.3**, we note that the predictions would have been different had we tested this assumption against the opposite, less natural assumption that label II players are commonly viewed as more sophisticated. Moreover, it is immediate from the experimental findings in the next section that subjects' behavior is consistent with

IA.3 and would reject the alternative. We also view assumption **IA.4** as natural, although it can be relaxed to allow for more noisy beliefs with minimal impact on our interpretation of the results. That said, it is impossible to guarantee that our assumptions, or small variants thereof, are the only ones consistent with the data in the universe of all conceivable assumptions. While we could document our results without such assumptions, the possibility to generate testable predictions and the interpretation of our results rely on the link that our assumptions establish between our model and the treatments. In that sense, our approach is subject to the nearly inescapable issues that characterize identification strategies in structural models.

As an illustration of alternative identification assumptions, suppose that instead of setting the cognitive bound to 0, the tutorial had an effect exclusively on subjects' beliefs and no impact on cognitive bounds. It is perhaps difficult to see why our tutorial would affect beliefs over behavior in such a manner instead of through the channel of cognition (notice that our assumption also leads to a difference in beliefs over behavior, but not over opponents' cognitive costs), but this assumption would also be consistent with the results presented in Section 3.5. We do not use such an approach because it is unclear to us why a tutorial would impact beliefs over behavior directly, rather than through the channels we describe. We believe that our assumption of the tutorial reducing the cost of reasoning to 0 is more plausible, however. The tutorial details the whole chain of reasoning. If we accept the weaker assumption that it is costly to think the game through in a game-theoretical manner, then the assumption that it becomes costless after the solution has been provided should only be a small step.[16]

Another alternative assumption to ours is that subjects are willing to put in more effort when facing high type rather than low type subjects, which would shift the subjects' cognitive bounds. This could

[16]Note as well that assumption **IA.1** allows for a shift in beliefs over behavior post-tutorial, but not over opponents' cognitive costs. Importantly, it allows for behavior to be driven by beliefs only.

be, for example, if the subjects feel particularly competitive with high type opponents. We cannot rule out that this occurs. Yet, we do not believe this to be the case, as subjects will not find out who they 'beat' or not, as no feedback is provided. It would therefore be surprising if the differential competitive aspect were a strong factor here. But we mention these alternatives to demonstrate that our approach is not immune to the issues common to such identification exercises, and that alternative assumptions can always be found to accommodate the observed behavior.

3.5 Results

Before examining the results of the individual experiments, we compare our general results to the findings of other level-k papers. Georganas et al. (2015) find that it is difficult to compare different types of level-k games as they find that levels of thinking can be uncorrelated or even flip across games. For this reason, we compare our results to Arad and Rubinstein (2012) who use the (cyclical) 11-20 game. They find that the number 20 is played by 6% of subjects, numbers 17, 18 and 19 are played by 74% of subjects and numbers from 11 to 16 are played by 20% of subjects. For our experiments, we find that 8% to 10% play 20, 50% to 60% play numbers 17 to 19 and 32% to 43% play 11 to 16 depending on the experiment. In light of the difference between the cyclical and our acyclical version of the game, these numbers appear comparable with those in Arad and Rubinstein (2012): the higher percentages of subjects who play numbers 11 to 16 are due to the fact that in our 11-20 game, there are no cycles and 11 is the level-∞ (and level-9 and higher) strategy, while in the 11-20 game in Arad and Rubinstein (2012), 11 is not level-∞ due to the cyclicality of the game. Our results for the percentages of levels played are also similar to those in Agranov et al. (2012) where 8% to 10% play level 0 (our outcome 20) and 42% to 77% play levels 1 to 3 (our outcomes 19, 18 and 17). The following sections will present the

experiment-specific results.

3.5.1 Experiment 1

In the following, we will combine the labels for the two classification criteria, and use the term 'label I' to refer indiscriminately to the 'math and sciences' or to the 'high score' subjects, and the term 'label II' to refer to the 'humanities' or 'low score' subjects.

3.5.1.1 Summary of AP's main results

AP's main findings on the baseline treatments can be summarized as follows:[17]

1. *Beliefs Effects:* For both the low and the high payoff treatments, under both classifications, the distribution of actions for label I subjects is *lower* in the homogeneous than in the heterogeneous treatments. The opposite is true for label II: the distribution of actions for label II subjects is *higher* in the homogeneous than in the heterogeneous treatments. Hence, these patterns are consistent with the assumption that both groups regard label I subjects as 'more sophisticated' than label II.

2. *Payoffs Effects:* For all configurations of beliefs, under both classifications, the distribution of actions in the 'low payoffs' treatments, $[X]$ – where $X = Hom, Het, HOB$ – first-order stochastically dominates the 'high payoffs' treatments, $[X+]$, for both label I and label II subjects. Hence, holding beliefs constant, the distribution of actions shifts towards higher level-k's when payoffs increase, which is consistent with Proposition 2.1. See Table 3.10 in the Appendix for the regression results.

[17] Other than the content of the next three bullet points, which summarize the experimental findings in AP, all other experimental results in this paper are new.

3. *Higher-Order Beliefs Effects:* Under both classifications, the distribution of actions for label I subjects is *lower* in the heterogeneous treatment [Het] than in the replacement treatment [HOB]. This suggests that label I subjects expect label II subjects to behave according to higher k's when they interact with label I, than when they play among themselves, and that label I subjects react to this. For label II subjects instead the distribution of actions in the [Het] and [HOB] treatments are essentially the same. Hence, higher-order beliefs effects are present, but they are 'one-sided', consistent with Proposition 3.1.

Under the assumption, which in fact emerges from the data, that both groups regard label I subjects as 'more sophisticated', the predictions of the EDR model are exactly those observed in the experiment, including the one-sidedness of the higher-order beliefs effects (cf. Alaoui and Penta (2016a)). In the rest of this section we discuss our novel experimental results.

3.5.1.2 Relaxing cognitive bounds – Experimental Results

In this subsection we discuss our findings for the post-tutorial treatments ([Tut], [AT-Hom] and [AT-Het]), which we administered to all eighty subjects from four of the six sessions (two for the endogenous and two for the exogenous classifications), each repeated twice.

Unsurprisingly, a high fraction of the subjects in treatment [Tut] (48% of label I and 55% of label II) chose 11, although one could have expected that an even higher fraction would have made that choice.

The empirical analysis conducted throughout is as follows. We use panel regressions, clustered at the individual level for all the analyses. We also check all of these comparisons for robustness with a Wilcoxon signed-rank test, and they generally confirm the results of the regressions discussed in this paper. For brevity, we omit the Wilcoxon signed rank tests' p-values from the main text but provide them in Table 3.16 in the Appendix and discuss those of note.

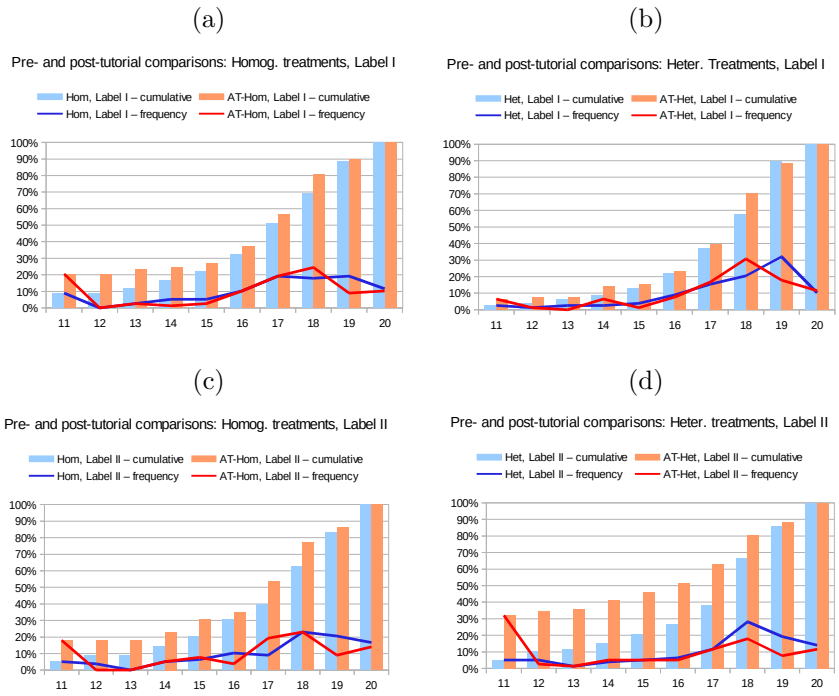


Figure 3.1: Pre- and post-tutorial comparisons, label I (top) and label II (bottom)

We first consider treatments pre- and post-tutorial. Comparing [Hom] to [AT-Hom], we observe that the distributions of actions shift to the left, with complete first order stochastic dominance (FOSD) of [Hom] to [AT-Hom] for both labels I and II (see Figure 3.1).[18] These results are supported by the regressions performed. For each label, we regress the chosen action on a dummy which takes value 1 if the treatment is [AT-Hom] and 0 if it is [Hom] (see Table 3.11). Consistent with the patterns observed in Figure 3.1, the estimated coefficient is negative (-0.71 for label I, and -0.85 for label II), statistically significant at the 5% level for label II and nearly significant at the 10% level for label I (the Wilcoxon signed-rank test is significant at the 10% level).[19] This means that the tutorial, going from [Hom] to [AT-Hom], induces an average decrease in the number chosen by subjects equal to 0.71 and 0.85 for the two labels. Similarly, when comparing [Het] to [AT-Het], we find relatively weak FOSD everywhere of [Het] over [AT-Het] for label I except at 19, and lack of significance for the estimated coefficient. For label II, there is stronger FOSD everywhere. The estimated coefficient takes value -1.92, and is statistically significant at the 1% level.

These results are all consistent with Proposition 2.3. The shift to

[18] While this is not our focus, we also check whether labels responded differently to treatments by conducting panel regressions with a label I dummy, a treatment dummy and an interaction term. The results are provided in Table 3.13. The coefficient of the interaction term is significant at the 1% level for the comparisons of the treatments from [Het] to [AT-Het] and [AT-Hom] to [AT-Het], and not the others. Note also that mechanically, if we compare the coefficients of this regression with the regressions that split the sample into label I and label II, the treatment effect on label II is of course identical to the one in the other regressions for label II. For label I, the sum of the treatment and interaction coefficients is identical to the regression results for label I in the other regressions.

[19] In all the regressions that follow, we control for the grouping of the exogenous and endogenous treatments. This control is never statistically significant at the 5% level and only once at the 10% level, and it has a marginal impact on the relevant coefficient compared to when it is omitted. For this reason, we do not discuss it below.

lower numbers for label II going from [Hom] to [AT-Hom] and [Het] to [AT-Het] indicates that the capacity, or cognitive bound, is binding for at least some of the subjects of that label when playing [Hom] and/or [Het]. Recall that we can make this inference because subjects face essentially the same opponent in [Hom] and [AT-Hom] (and [Het] to [AT-Het]), and so the tutorial may affect their behavior only if the previous understanding was binding.[20] Similarly, the shift to lower numbers for label I when going from [Hom] to [AT-Hom] also indicates that the cognitive bound is binding in [Hom] when playing against their own type. Moreover, the lack of significance of the coefficient when comparing [Het] to [At-Het] for label I does not allow us to conclude that their cognitive bound was binding in [Het]. This is also in line with our predictions, given our maintained assumption (IA.3) that label II is commonly believed to be less sophisticated than label I.

Interestingly, Figures 3.1 and 3.2 show that players jump to 11 in the post-tutorial treatments much more frequently when their opponent is (weakly) more sophisticated. From within the EDR model, this suggests that subjects believe the more sophisticated players to be very sophisticated, and capable of playing according to the highest level of reasoning. Their behavior against the less sophisticated opponent provides a further indication that it is beliefs that drive their behavior against them, and not their cognitive bound. It also suggests that the tutorial served its intended purpose of lowering the costs of cognition significantly. Within the context of the EDR model, it is a direct implication of identification assumption IA.1.[21]

We now analyze differences between homogeneous and heteroge-

[20] We note that the fact that subjects face the same population of opponents is a feature of the experiment. Its formal counterpart within the language of the EDR model is provided by identification assumption IA.2.

[21] A small caveat could be the wording of the tutorial as it mentions "game theory" and "rationality". Subjects might not completely understand these terms. However, the reaction of subjects to the tutorial suggests that this was not an issue.

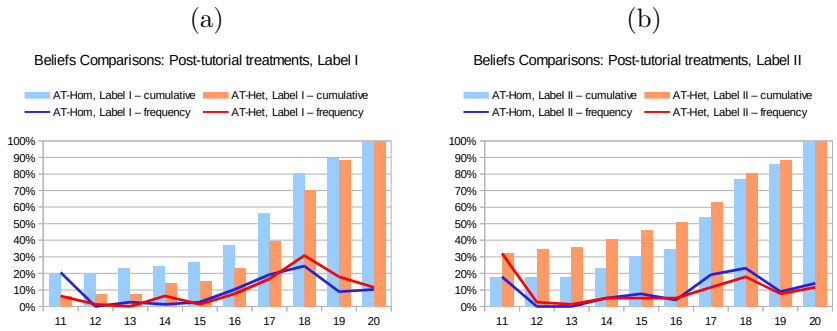


Figure 3.2: Beliefs comparisons: post tutorial treatments, label I (left) and label II (right)

neous post-tutorial treatments. Comparing [AT-Hom] to [AT-Het] for each label (see Figure 3.2), in the case of label I, there is pronounced FOSD of [AT-Het] over [AT-Hom] everywhere, and the estimated coefficient of the regression is 1.06 and statistically significant at the 1% level. In the case of label II, instead, the effect is reversed: There is strong FOSD of [AT-Hom] over [AT-Het] everywhere, and the estimated coefficient is -1.14, also statistically significant at the 1% level. Here as well, the direction of the results is fully consistent with our predictions (Proposition 3.2). Notice that the tutorial should not affect the direction of the comparative statics, because the main relevant factor is not the subjects' own understanding of the game, but rather their beliefs over their opponents' understanding.

Lastly, since we expect more label II subjects with binding cognitive bounds in treatment [Het] than in treatment [Hom] (and in [Het+] than in [Hom+]) – and the opposite for label I subjects – the model (Proposition 4) predicts that we should observe a (weakly) greater shift in behavior from [Het] to [AT-Het], than from [Hom] to [AT-Hom] for label II subjects – and the opposite for label I. This is consistent with the empirical findings in Figure 3.1, and also in line

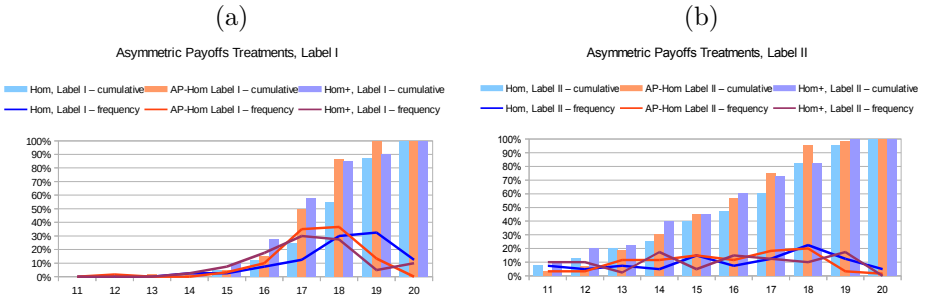


Figure 3.3: Asymmetric payoffs treatments, label I (left) and label II (right).

with the estimated OLS coefficients: the estimated coefficients for label II are -1.92 (significant at 1%) going from [Het] to [AT-Het] and -0.85 (significant at 5%) going from [Hom] to [AT-Hom]; for label I subjects instead, the coefficients are -0.7 (significant at (nearly) 10%) going from [Hom] to [AT-Hom], and -0.32 (not significant) going from [Het] to [AT-Het]. These comparisons are consistent with the model's predictions.

3.5.1.3 Reasoning about opponents' incentives – Experimental Results

We now discuss the empirical findings for the asymmetric payoff treatments [AP-Hom] and [AP-Het], which were administered after the baseline treatments to all forty subjects from two sessions (one exogenous and one endogenous), each treatment repeated three times. All of the results of the regressions discussed in this subsection are provided in Table 3.12.

Comparing [Hom], [AP-Hom] and [Hom+] for label I, there is a nearly complete FOSD relationship of [Hom] over [AP-Hom] (and of [Hom] over [Hom+]), but the relationship between [AP-Hom] and [Hom+] is less clearly defined (see Figure 3.3). This is consistent with

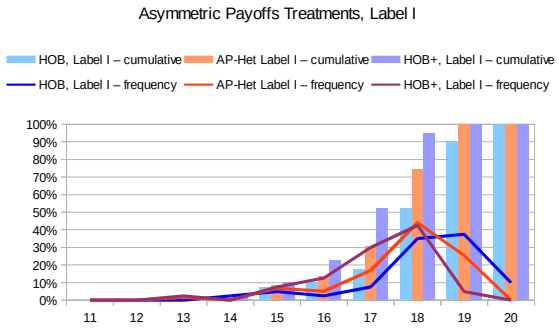
the regressions, for which the estimated coefficient when going from [Hom] to [AP-Hom] is -0.74, and is statistically significant at the 1% level, while it is not significant when going from [AP-Hom] to [Hom+]. For label II, the comparisons of [Hom], [AP-Hom] and [Hom+] are ambiguous, and neither of the regressions comparing [Hom] to [AP-Hom] or [AP-Hom] to [Hom+] lead to significance. These results are jointly in line with Proposition 2.2(i) (recall that those predictions are in terms of weak orderings). Moreover, from within the EDR model they indicate that, for label I, increasing only the individual's own incentives, without changing either the opponents' incentives or their beliefs over their opponents, leads to them playing according to higher rounds of reasoning. In other words, the change in incentives appear to have led some label I subjects to increase their cognitive capacity. For label II subjects, instead, the increase in incentives has not had a noticeable impact.

When considering the [AP-Het] treatment, the natural comparison is not [Het], but rather [HOB]; see Proposition 2.2(ii). This is because the only difference between [HOB] and [AP-Het] is in the incentives of the subject, while their opponents are identical in the game they play. With [Het], instead, there is also a difference in the opponents, in that the opponents' opponent is of a different label. Comparing [HOB], [AP-Het] and [HOB+] for labels I and II separately, we see clear FOSD relationships nearly everywhere of [HOB] to [AP-Het] to [HOB+] for label I (see Figure 3.4). The relationship between the curves is instead more ambiguous for label II. The regressions for these comparisons lead to statistical significance at the 1% level for label I, but they are not significant for label II. These results are consistent with Proposition 2.2(ii).

Lastly we compare [AP-Hom] to [AP-Het] for both labels, and find a FOSD relationship of [AP-Het] to [AP-Hom] (see Figure 3.5). The coefficient estimated in the regressions, however, is not significant for label II, while it is significant at the 5% level for label I.

The model, and specifically Proposition 3.4, which predicts a shift from [HOB] to [AP-Het] of at most 1, also seems consistent with

(a)



(b)

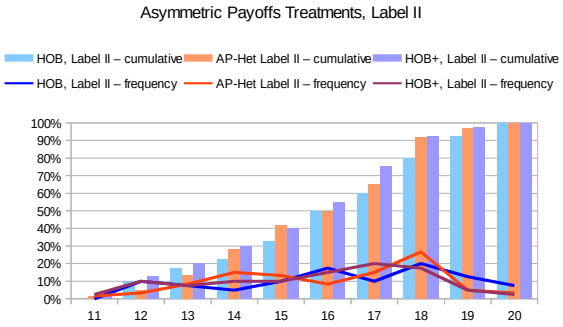


Figure 3.4: Asymmetric payoffs treatments, with double replacement. Label I (top) and label II (bottom).

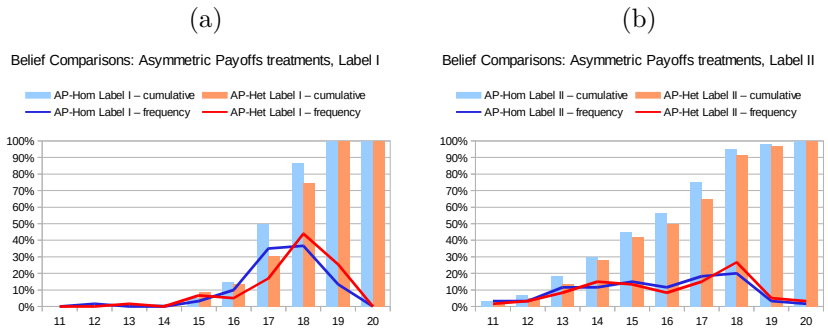


Figure 3.5: Beliefs comparisons: asymmetric payoffs treatments, label I (left) and label II (right).

the small shift in distribution from [HOB] to [AP-Het], and with the estimates of the OLS coefficient (-0.5, significant at the 5% level). In this case, and consistently with the theory, the movement from [HOB] to [HOB+] for label I is mainly due to the increase in the opponents' payoffs, and not solely to the agent's own incentives. In light of the complexity of these treatments and the difficulty of the instructions, both discussed in Section 3.3.1.3, these results are remarkably consistent with the predictions of the EDR model.

Robustness. As discussed in Section 3.3.1.3, we have conducted an additional robustness experiment that also contains treatments with replacement both for the [Hom-Rep] benchmark and the replacement treatment [AP-Hom]. It also includes the [Hom] treatment, for comparability. This serves to address the possible concern that the replacement method might remove social preferences. A Wilcoxon signed-rank test (selected due to the dependence across the treatments) could not reject the null that behavior in [Hom] and [Hom-Rep] was the same for the whole sample and the label I and label II subsamples. This suggests that there is no confound of social preferences in the [Hom] treatment compared to a replacement treatment and

that we can therefore use the results of Experiment 1 for our analysis.

A second possible concern is that order effects might be the reason for increases in the levels played. We deal with this in three ways (in both the original Experiment 1 and the robustness experiment). First, the order of treatments was randomized across sessions so that order effects should not play a role. Second, all of our reported regression results (for all experiments) were estimated using panel regressions that took into account that the same treatment was asked multiple times. Third, we test for order effects using Wilcoxon signed-rank tests. We find no order effects for repeated treatments other than for [Hom+] in those experiments with tutorial treatments (p-value=0.055) and for [HOB] for those experiments with asymmetric payoffs treatments (p-value=0.064). For those treatments where no order effects were observed, p-values range from 0.15 to 1 with the majority above 0.3. This suggests that order effects are unlikely to drive the results.

3.5.2 Experiment 2: Results

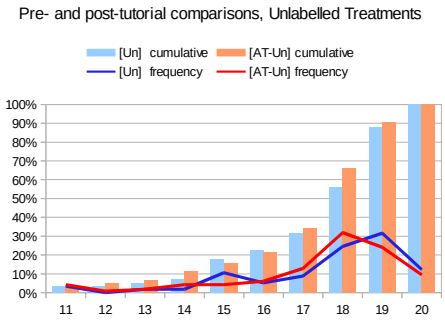


Figure 3.6: Pre- and post-tutorial, unlabeled treatments.

We first compare the results for treatments [Un] and [AT-Un]. Graphically, the cumulative distribution for [Un] stochastically dominates [AT-Un] everywhere except at 15 and 16 (Figure 3.6) but the

difference between the distributions is slight. We find that the estimated coefficient in the regression is not significant but has the expected sign (see Table 3.14 for the regressions discussed here). The relationship between [Un] and [AT-Un] is consistent with Proposition 2.5, which predicts weak stochastic dominance. We also note that the Wilcoxon signed-rank test comparing the distributions is significant at the 5% level. Put together, we cannot conclude from these results to which extent subjects' behavior in the [Un] treatment is driven by their own cognitive capacity or by their beliefs about their opponents.

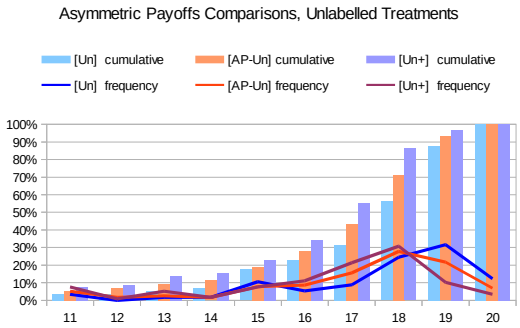


Figure 3.7: Asymmetric payoffs comparisons, unlabeled treatments.

Comparing the distributions of treatments [Un], [AP-Un] and [Un+], we observe that [Un] first-order stochastically dominates [AP-Un] everywhere, which itself stochastically dominates [Un+] everywhere (see Figure 3.7). Consistent with these results, the estimated coefficients are statistically significant for the regressions comparing [Un] to [AP-Un], [AP-Un] to [Un+] and [Un] to [Un+] at the 10% level (p-value 0.054), 5% level and 1% level, respectively. These findings indicate that subjects play according to lower sophistication in [Un] than [AP-Un] than [Un+]; this is consistent with Proposition 2.4.

The difference between [Un] and [AP-Un] is in the incentives, holding constant beliefs and higher-order beliefs over the distributions of opponents. Hence, playing according to higher sophistication in [AP-

Un] than in [Un] is an indication of the cognitive bound increasing. In the comparison between [AP-Un] and [Un+] instead, agents have the same incentives, and hence the difference between the two treatments is due to subjects' beliefs over the opponents. Specifically, since [AP-Un] and [Un+] differ in the opponents' incentives to reason, the fact that behavior is markedly different in these two treatments (the OLS coefficient is -0.55, with p-value of 0.023) is a clear indication that subjects take into account their opponents' incentives to reason, when they form beliefs over their behavior.

Lastly, Wilcoxon signed-rank tests that examine whether different instances of the same treatments can be differentiated from each other suggest that there are no order effects (p-values range from 0.35 to 0.88).

3.5.3 Experiment 3: Results

Recall that one of the main objectives of Experiment 3 is to test whether our results comparing pre- and post- tutorial treatments in the earlier experiments could have been driven by the tutorial inducing level-k reasoning rather than by reducing the cognitive costs. Since, in Experiment 3, all subjects were already primed to think according to level-k when playing the first treatment [E3-Un], changes observed between the pre- ([E3-Un]) and post-tutorial responses ([E3-AT-Un]) should not stem from priming but from changes in their cognitive levels.

We observe that [E3-Un] first order stochastically dominates [E3-AT-Un] everywhere (see Figure 3.8) and the regression coefficient is significant at 10% (see Table 3.15). This result is consistent with Proposition 2.5, and shows that because subjects played according to a higher level after receiving the tutorial the cognitive constraint in [E3-Un] must have been binding for some subjects. We also find that subjects played according to higher levels going from [E3-Un] to [E3-AP-Un] and from [E3-Un] to [E3-Un+] (significant at 5% and 1%, respectively). This suggests that the higher incentive in payoffs

led to more rounds of reasoning (see Figure 3.9) and is consistent with Proposition 2.4. The results for Experiment 3 are all consistent with the predictions of our model, and with the results of the other experiments. It does not seem, therefore, that the potential level-k priming effect of the tutorial is driving those findings.

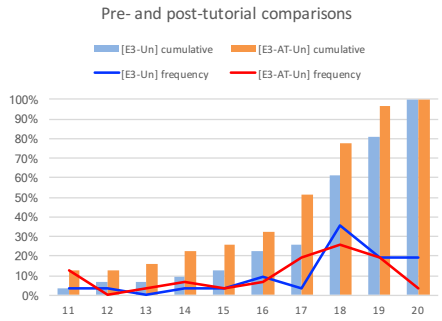


Figure 3.8: Pre- and post-tutorial, unlabeled treatments, Experiment 3.

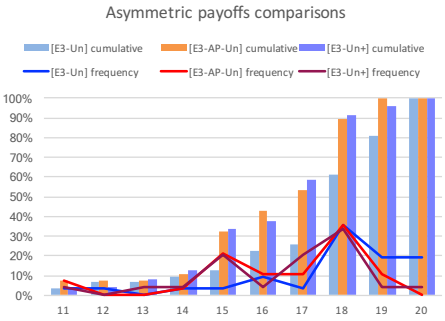


Figure 3.9: Asymmetric payoffs comparisons, unlabeled treatments, Experiment 3.

The experimental results show that the way that subjects react

to changes in the label of the opponent, and in changes of the label of their opponent's opponent, is entirely consistent with the higher order beliefs effects generated by the EDR model (namely, that they are one-sided, and that they interact in complex ways with other changes in the environments, such as asymmetric changes in the payoff structure).

While our approach is *as if*, and we do not take a stance on the actual deliberation process of the agent, one could wonder whether subjects are *actually* capable of performing higher order reasoning (see Kets (2017) and Heifetz and Kets (2018) for models which formalize the idea that players need not have well-formed higher order beliefs). To gain a more direct answer to this question, we administered a short version of the Theory of Mind test (TOM hereafter; see Stiller and Dunbar (2007), Liddle and Nettle (2006)) at the end of the experiment. This test is aimed specifically at testing whether a subject can place themselves in the mind of another person. Subjects are given multiple short stories to read about the interaction between fictional characters. In these stories, the main character is thinking about the motivation behind statements or actions of others. After each story, subjects are asked to complete a series of questions about the story. To rule out that bad performance in the test is based on lack of attention or memory, the test contains a factual part aimed at testing how well subjects remember the story. In the mind part, subjects have to answer questions about others' reasoning process (and others' reasoning process about others' reasoning up to several levels). Bad performance in both parts suggests that the subject did not remember the story correctly while good performance in the factual part coupled with bad performance in the mind part suggests that the subject is able to remember the story but unable to place themselves in the mind of someone else. Test results show that more than 70% of subjects answered more than 50% of the TOM questions correctly (see the distribution of TOM scores in Figure 3.10). The results are not driven by the results of the factual part which can be seen in Figure 3.11 in the Appendix. Most of the TOM questions

require higher order beliefs of multiple levels. Only one subject was unable to correctly answer any question that requires one level of thinking. This suggests that the majority of test subjects is capable of reasoning about others' reasoning.

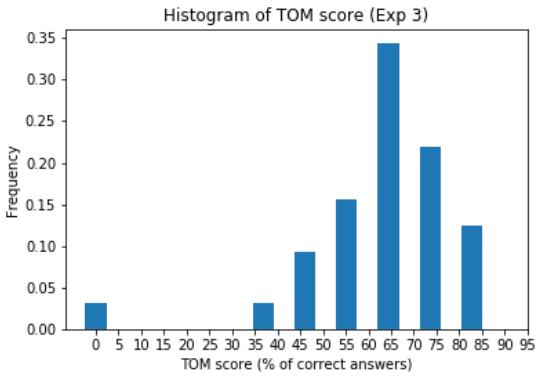


Figure 3.10: Distribution of total TOM scores.

In addition to the TOM, subjects had to complete a test of reasoning before starting the experiment. This test contained the muddy faces (also known as dirty faces) game which examines the 'level of iterated rationality' (Weber (2001)). On average, subjects scored nearly 85% in this test, suggesting that they are capable of reasoning iteratively which is a prerequisite of level-k reasoning. Together, these results suggest that the experimental subjects are indeed capable of reasoning about others' reasoning and that we can thus use the results from the replacement method.

To test for the potentiality that test results are driven by order effects, we conducted Wilcoxon signed-rank tests that examine whether different instances of the same treatments can be differentiated from each other. The resulting p-values suggest that there is only one order effect present in the experiment; [E3-AT-Un] with a p-value equal to 0.057 (p-values for the other treatments range from 0.41 to 0.49). The results are thus unlikely to have been driven by order effects and,

Section Nr.	Experiment Nr.	Subject groupings	Treatments	Propositions tested	Comments
3.5.1	1	Labeled	[Hom], [Het],[HOB], [Hom+], [Het+], [HOB+], [Tut], [AT-Hom], [AT-Het], [AP-Hom], [AP-Het]	2.1, 2.2, 2.3, 3.1, 3.2, 3.3, 3.4, 4	Results robust to order effects & to replacing [Hom] with [Hom-Rep].
3.5.2	2	Unlabeled	[Un], [Un+], [Tut-Un], [AT-Un], [AP-Un]	2.4, 2.5	Results robust to order effects.
3.5.3	3	Unlabeled, semi-tutorial	[E3-Un], [E3-Un+], [E3-Tut-Un], [E3-AT-Un], [E3-AP-Un]	2.4, 2.5	Results robust to semi-tutorial and to order effects.

Table 3.9: Summary of results over all experiments.

as in the other experiments, we use panel regressions to take into account that treatments were administered multiple times.

Table 3.9 summarizes the main results of Experiments 1, 2 and 3.

3.5.4 Individual Effects

The experimental results sections above analyze behavior at the aggregate level. One possible concern with this approach is that behavior within an individual might not be consistent and that this might be averaged out. In this section, we examine individual-level results which can shed light on this concern. We find that most subjects are highly consistent in their behavior across treatments.

To analyze individual level behavior across treatments, we created a score of violations against the theory. For example, the theory predicts that the level played in [Hom] is weakly lower than that in [Hom+]. A violation would then be if the subject played a strictly higher level in [Hom] (similarly for [Un] and [Un+] in experiments 2 and 3). For all individuals across all experiments, we count the number of violations across all comparisons of treatments where the theory makes a clear prediction. Figures in Section 3.7.4.2 of the Appendix show that for each experiment, the number of violations

is relatively low. For example, for Experiment 2, more than 60% of subjects had zero violations of the theory and another 15% had only one violation of the theory in the averaged treatment violations score. These results suggest that subjects are highly consistent in their behavior. The results for each experiment individually can be seen in the Appendix.[22]

A second potential issue with looking at aggregate results is that some individuals might shift their behavior a lot, driving the observed average effects. In order to examine this possibility, we calculated the number of levels that each subject shifts for all comparisons in the regressions. We then plot the distribution of these shifts in levels. In the figures in section 3.7.4.3 of the appendix, it can be seen that most of the subjects shift levels moderately and for most comparisons a very low percentage of subjects shifts by many levels.[23] Due to their very low frequency, it is unlikely that the results are driven by these individuals.

These figures also show that the shift in levels of reasoning is larger for the tutorial than the payoff treatments. This suggests that the tutorial is successful at removing the cognitive constraint so that levels of play are determined by beliefs only. Finally, these figures allow us to examine by how many levels label I subjects shifted from [HOB] to [AP-Het]. Figure 3.17 shows that 80% of label I subjects shifted by one level or less, which supports Proposition 3.4.

[22]Violations can also be calculated based on instances of each treatment i.e. comparing the level played in the first instance of [Hom] to the first instance of [Hom+] and the second instance of [Hom] to the second instance of [Hom+] and so forth. Figures for these violation scores are available upon request.

[23]Due to the large number of figures, we only show the results for Experiment 3 and one figure from Experiment 1 that is mentioned above. All other figures are available upon request.

3.6 Concluding Remarks

Individuals may reason about others' strategic reasoning in a way that is more nuanced than has been typically considered by the existing literature. In particular, it is not clear whether subjects' choices are constrained by their own cognitive limitations or rather by their beliefs about their opponents' limitations. It is also unclear whether subjects would take into account how changing the opponents' stakes may change their depth of reasoning, and hence their behavior. In this paper we provided several experiments designed to address these questions.

Our findings indicate that subjects do indeed reason about their opponents' reasoning process, and that they form beliefs not only about their opponents' sophistication, but also account for the change of this sophistication with the incentives to reason. We also find that, while beliefs play a clear role in subjects' behavior, the cognitive bounds of a significant fraction of subjects are binding and determine their behavior when facing opponents they view as more sophisticated. These results suggest that, in general, level-k behavior should not be taken as driven either by cognitive limits alone or beliefs alone. In some settings it is a function of both, and depends on the complex interaction between cognitive bounds, beliefs about the opponent's cognitive abilities, and reasoning about others' reasoning. We also find that the EDR framework of Alaoui and Penta (2016a) is a useful tool for analyzing and understanding this interaction, and that the results are overall consistent with its predictions.

While we have focused our analysis on a specific game and setting, our experimental design can be applied to a number of other contexts. It is based on two key ideas, which we have referred to as the *replacement* and the *tutorial* methods. These methods are to a large extent independent of the features of the underlying game, and are thus portable and easily implemented in other settings as well.

The *replacement method* consists of replacing the opponent's opponent and controls for higher-order beliefs effects. Controlling subjects'

beliefs hierarchies is a well-known difficulty in designing game theoretic experiments, particularly if they are aimed at isolating the effects of beliefs manipulations or at identifying subjects' higher order reasoning.[24] The precise wording of the treatments based on the replacement method is designed to pin down the entire hierarchy of beliefs, thereby addressing this challenge.[25]

The *replacement method* was used in several of our treatments. The basic idea, however, has much broader applicability. It can be applied to essentially any setting (level-k or not) to disentangle *direct* and *interactive* effects involved in general comparative statics exercises. For instance, it can be used to separate subjects' own preferences to adhere to a social norm from their beliefs about the consequences faced when deviating from it. The method can also be further adapted to more complex settings, such as the interesting experiment by Proto et al. (2019), to assess for instance to what extent the higher levels of cooperation observed in the high-IQ group are due to individuals' own cognitive abilities, or to the high-IQ environment.

The *tutorial method* instead was designed to address the cognition-beliefs dichotomy. It consists in studying how subjects' behavior is

[24] We stress that the importance of the replacement method is to disentangle various higher order beliefs effects by progressively changing the orders of beliefs one-by-one, keeping all higher order beliefs constant from one design to the next. In this sense, while the experiment in Agranov et al. (2012) contains treatments analogous to our [Hom] and [HOB] treatments, it still cannot be considered as a full implementation of the method we are describing, since comparing [Hom] and [HOB] directly does not enable us to disentangle first-order effects from higher-orders. We also note, as we discuss next, that the replacement method can be applied to beliefs as well as payoffs, as we did in the AP-treatments.

[25] In a recent paper, Kneeland (2015) develops the idea of 'ring games', which has a similar objective. One important difference between ring games and the replacement method is that, by having player 1 have a dominant strategy, Kneeland's ring games allow for an exact identification of different orders of beliefs in rationality, which our replacement method does not. In contrast, the advantage of replacement treatments is that it allows to investigate higher order beliefs effects (albeit partially identified) in arbitrary games, without altering the underlying payoff structure, in a way that is easy to implement in the lab.

affected by receiving a tutorial which contains non-factual information about the strategic structure of the game. As long as subjects face the same opponents before and after having received this tutorial (an idea related to the replacement method discussed above), then their behavior should change if and only if the tutorial has made them understand something they deem useful. This has enabled us to assess whether subjects' understanding was somehow binding, relative to the information provided by the tutorial, or whether their action was rather driven by their beliefs, whatever understanding of the game they had before or after the tutorial. This idea can be applied independently of whether the underlying reasoning takes the form of level-k thinking. For instance, one could imagine a game with multiple equilibria (e.g., stag hunt, or other coordination games in which subjects may resort to different, non level-k, reasoning processes), and have the tutorial simply explain the properties of the different equilibria (such as risk-dominance and efficiency in stag-hunt). The exercise would go through in that case essentially unchanged, to assess the extent to which subjects' understanding before the tutorial was binding or not.

Similar to the replacement method, the tutorial method is portable to other settings, and is especially suited to understanding the cognition-beliefs dichotomy for different forms of reasoning. It need not only apply to level-k reasoning or to games of initial response. Future research can therefore make use of both these methods in settings that are very different from the one focused on in this paper.

Acknowledgements:
Some of the material presented in this paper circulated in an old working paper of ours, entitled 'Level-k Reasoning and Incentives' (Alaoui and Penta (2012)).
We are indebted to Vincent Crawford, Robin Dunbar, Christian Fons-Rosen, Nagore Iriberri, Christian Michel, Rosemarie Nagel, Marciano Siniscalchi, Alessandro Tarozzi and two anonymous referees for the generous and very helpful comments on the early drafts of this project. We also thank several seminar and conferences audiences where these results were presented. The authors acknowledge financial support from the Spanish Ministerio de Economía y Competitividad, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0563). Larbi Alaoui acknowledges financial support from the Ministerio de Economía y Competitividad (grant number PGC2018-098949-B-I00) and the Ministerio de Ciencia e Innovación (Beca Ramon y Cajal RYC-2016-21127). Antonio Penta acknowledges financial support from the ERC Starting Grant #759424.

3.7 Appendix

3.7.1 Proofs

Proof of Proposition 1. For point 1, suppose that $c_j^i, c_i^{ij} \in C^+(c_i)$:

- Let v_i denote i's value of reasoning in the low payoff game, and notice that assumptions (i)-(v) in pp. 128-129 imply that $v_i = v_j^i = v_{i(j)}$ in that game. Hence, if $c_j^i, c_i^{ij} \in C^+(c_i)$, we have $\mathcal{K}(c_j^i, v_j) \geq \mathcal{K}(c_i, v_i) = \hat{k}_i$ and $\mathcal{K}(c_i^{ij}, v_{i(j)}) \geq \mathcal{K}(c_i, v_i) = \hat{k}_i$. The former inequality implies (by eq. (3.3)) that $\hat{k}_j^i = \hat{k}_i - 1$, which together with the latter inequality and (3.4) implies $\hat{k}_i^{ij} = \hat{k}_i - 2$. By (3.5), it follows that $k_j^i = \hat{k}_i - 1$, and hence $k_i = k_j^i + 1 = \hat{k}_i$.

- Let v_i' denote i's value of reasoning for the high x_i in the asymmetric payoff game. By assumptions (i)-(v) in pp. 128-129, v_j^i and $v_{i(j)}$ remain the same as in the low payoff case, whereas $v_i'(k) \geq v_i(k)$ for every k. Letting $\hat{k}_i' := \mathcal{K}(c_i, v_i')$ and k_i' denote, respectively, i's cognitive bound and behavioral level in the asymmetric payoff game, it follows that $\hat{k}_i' \geq \hat{k}_i$. Equations (3.3)-(3.5) then immediately imply that $\hat{k}_i^{ij}$ and k_i^j weakly increase from the low-payoff to the asymmetric payoff game, and hence $k_i' \geq k_i$.

- Let $\tilde{v}_i$, $\tilde{v}_j$ and $\tilde{v}_{j(i)}$ denote the value of reasoning in the high payoff game for i ,j and j's opponent, respectively. By assumptions (i)-(v) in pp. 128-129, $\tilde{v}_i = v_i'$ and $\tilde{v}_i = \tilde{v}_j^i = \tilde{v}_{i(j)}$. Hence, if $c_j^i, c_i^{ij} \in C^+(c_i)$, we have $\mathcal{K}(c_j^i, \tilde{v}_j) \geq \mathcal{K}(c_i, \tilde{v}_i) = \mathcal{K}(c_i, v_i') = \hat{k}_i'$ and $\mathcal{K}(c_i^{ij}, \tilde{v}_{i(j)}) \geq \mathcal{K}(c_i, \tilde{v}_i) = \mathcal{K}(c_i, v_i') = \hat{k}_i'$. Letting k_i'' and $\hat{k}_i''$ denote i's behavioral level and cognitive bound in the high payoff game, the same arguments as in the low payoff game at this point imply that $k_i'' = \hat{k}_i'' = \hat{k}_i'$, which in turn implies that $k_i'' > k_i'$ only if $k_i' < \hat{k}_i'$.

For point 2, first consider the case $c_j^i \in C^-(c_i)$: Let v_i denote i's value of reasoning in the low payoff game, and notice that assumption (iii)-(v) in pp. 128-129 implies that $v_i = v_j^i = v_{i(j)}$ in that game. Hence, if $c_j^i \in C^-(c_i)$, we have $\mathcal{K}(c_j^i, v_j) \leq \mathcal{K}(c_i, v_i) = \hat{k}_i$.

- First note that, by eq. (3.3)-(3.5) and the definition of k_i, it is always the case that $k_i \leq \hat{k}_i$. The inquality can be strict only if $\mathcal{K}(c_j^i, v_j) < \hat{k}_i - 1$, which is possible (though not necessary) if $c_j^i \in C^-(c_i)$.

- Let v_i' denote i's value of reasoning for the high x_i in the asymmetric payoff game. By assumptions (i)-(v) in pp. 128-129, v_j^i and $v_{i(j)}$ remain the same as in the low payoff case, whereas $v_i'(k) \geq v_i(k)$ for every k. Letting $\hat{k}_i' := \mathcal{K}(c_i, v_i')$ and k_i' denote, respectively, i's cognitive bound and behavioral level in the asymmetric payoff game, it follows that $\hat{k}_i' \geq \hat{k}_i$. Since in the low payoff game we already had $\hat{k}_j^i \leq \hat{k}_i - 1$, and hence (by 3.4) $\hat{k}_i^{ij} \leq \hat{k}_i - 2$, it follows that the (weak) increase in i's cognitive bound does not affect either $\hat{k}_j^i$ or $\hat{k}_i^{ij}$. If in the low payoff game it was the case that $\hat{k}_j^i < \hat{k}_i - 1$ and $\hat{k}_i^{ij} < \hat{k}_i - 2$, and hence $k_j^i < \hat{k}_i - 1$, and hence $k_i = k_j^i + 1 < \hat{k}_i$, then we would have $k_i' = k_i$. It follows that if $k_i' > k_i$ (which could be the case if $\hat{k}_i^{ij} = \hat{k}_i - 2$, and $k_j^i = \hat{k}_i - 1$), then it must have been the case that k_i and $\hat{k}_i$ were equal in the low payoff game.

- The argument for the high payoff game is the same as the corresponding one for point 1 above.

Now consider the case, also in point 2, in which $c_j^i \in C^+(c_i)$ and $c_i^{ij} \in C^-(c_i)$:

- Again, note that, by eq. (3.3)-(3.5) and the definition of k_i, it is always the case that $k_i \leq \hat{k}_i$. Now, let v_i denote i's value of reasoning in the low payoff game, and notice that assumption (iii)-(v) in pp. 128-129 implies that $v_i = v_j^i = v_{i(j)}$ in that game. Hence, if $c_j^i \in C^+(c_i)$, we have $\mathcal{K}(c_j^i, v_j) \geq \mathcal{K}(c_i, v_i) = \hat{k}_i$. The former inequality implies (by (3.3)) that $\hat{k}_j^i = \hat{k}_i - 1$. Given this, $k_i < \hat{k}_i$ is possible only if $\mathcal{K}(c_i^{ij}, v_{i(j)}) < \hat{k}_i - 1$, which is possible (though not necessary) if $c_i^{ij} \in C^-(c_i)$.

  Given this, the arguments for the asymmetric and high payoff games are the same as those we considered above for the other case of point 2.

For point 3, let v_i, v_i' and $\tilde{v}_i$ denote, respectively, i's value of reasoning in the low, asymmetric, and high payoff games. As in point 1, the maintained assumptions imply that $v_i' = \tilde{v}_i$ and $v_i'(k) \geq v_i(k)$ for all k. Also, for any $w_i \in \{v_i, v_i'\}$ $\hat{k}_i' := \mathcal{K}(c_i', w_i) \geq \mathcal{K}(c_i, w_i) =: \hat{k}_i$ if $c_i' \in C^+(c_i)$ – that is, for any payoff configuration, i's cognitive bound always weakly increases as cost c_i is replaced with a lower cost c_i'. Since equations (by 3.3)-(3.5) are all (weakly) increasing in i's cognitive bound, and i's behavioral level is increasing in k_j^i, if follows that also the behavioral level (weakly) increases if c_i is replaced with a lower cost c_i'. □

Proof of Proposition 2. All results follow directly from Proposition 1 and IA1-2, noting that a higher behavioral level translates to lower numbers chosen in the acyclic 11-20 game:

1. Points 1 and 2 of Proposition 1 imply that, for any c_i, c_j^i and c_i^{ij} – which by IA-2 remain constant between X and $X+$ for any $X \in$ {Hom, Het, HOB} – the behavioral level weakly increases for any i from the low payoff to the high payoff game. Hence, for each $l \in \{I, II\}$: $F_X^l \succsim F_{X+}^l$ for all $X \in$ {Hom, Het, HOB}.

2. Points 1 and 2 of Proposition 1 imply that, for any c_i, c_j^i and c_i^{ij} – which, by IA-2, remain constant between [Hom], [AP-Hom] and [Hom+], and between [HOB], [AP-Het] and [HOB+] – the behavioral level weakly increases for any i from the low payoff to the asymmetric payoff game, and from the asymmetric to the high payoff game. The former increase is strict only if $\hat{k}_i = k_i$ in the low payoff game for some i. Hence, for each $l \in \{I, II\}$: (i) $F_{Hom}^l \succsim F_{AP-Hom}^l \succsim F_{Hom+}^l$, with $F_{Hom} \succ F_{AP-Hom}^l$ only if $\hat{k}_i$ was binding in [Hom] for some i; and (ii) $F_{HOB}^l \succsim F_{AP-Het} \succsim F_{HOB+}^l$, with $F_{HOB}^l \succ F_{AP-Het}$ only if $\hat{k}_i$ was binding in [HOB] for some i.

3. Point 3 of Proposition 1 implies that, for any c_j^i and c_i^{ij} – which by IA-2 remain constant between X and [AT-X] for any $X \in$ {Hom, Het} – the behavioral level weakly increases for any i if their costs are lowered from c_i to some $c_i' \in C^+(c_i)$, a condition which is satisfied in the post-tutorial treatment under assumptions IA.1, and strictly so

only if $\hat{k}_i = k_i$ in the first place. It follows that for each $l \in \{I, II\}$: $F^l_{Hom} \succsim F^l_{AT-Hom}$, and $F^l_{Het} \succsim F^l_{AT-Het}$, each strictly only if $\hat{k}_i$ was binding for some i in [Hom] and [Het], respectively.

4. Points 1 and 2 of Proposition 1 imply that, for any c_i, c^i_j and c^{ij}_i – which, by IA-2, remain constant between [Un], [AP-Un] and [Un+] in both experiments 2 and 3 – the behavioral level weakly increases for any i from the low payoff to the asymmetric payoff game, and from the asymmetric to the high payoff game. The former increase is strict only if $\hat{k}_i = k_i$ in the low payoff game for some i. Hence: (i) $F^*_{Un} \succsim F^*_{AP-Un} \succsim F^*_{Un+}$, with $F^*_{Un} \succ F^*_{AP-Un}$ only if $\hat{k}_i$ was binding in [Un] for some i; and (ii) $F^*_{E3-Un} \succsim F^*_{E3-AP-Un} \succsim F^*_{E3-Un+}$, with $F^*_{Un} \succ F^*_{E3-AP-Un}$ only if $\hat{k}_i$ was binding in [E3-Un] for some i.

5. Point 3 of Proposition 1 implies that, for any c^i_j and c^{ij}_i – which by IA-2 remain constant between Un and AT-Un in both experiments 2 and 3 – the behavioral level weakly increases for any i if their costs are lowered from c_i to some $c'_i \in C^+(c_i)$, a condition which is satisfied in the post-tutorial treatment under assumptions IA.1, and strictly so only if $\hat{k}_i = k_i$ in the first place. It follows that $F^*_{Un} \succsim F^*_{AT-Un}$ and $F^*_{E3-Un} \succsim F^*_{E3-AT-Un}$, each strictly only if $\hat{k}_i$ was binding for some i in [Un] and [E3-Un], respectively.

□

Proof of Proposition 3. Under the maintained assumptions (i)-(v) of the EDR model, plus identification assumptions IA.1-4, the cost of reasoning c_i and the value of reasoning v_i, v^i_j and $v^i_{j(i)}$ remain constant across all treatments within the same point of the proposition. Hence, also $\hat{k}_i$ does not change within each point of the proposition. The only things that change, within each point, are thus beliefs c^i_j and higher order beliefs c^{ij}_i. For each point we describe what these changes are and how they impact behavior across treatments:

1. IA.2 implies that $c^{i,[Het]}_j = c^{i,[HOB]}_j$ and $c^{ij,[Het]}_i = c^{ij,[Hom]}_i$ for any i of any label.

(a) For parts (i) and (iii), IA.3 implies that for any i with label I, we have $c_j^{i,[Het]} \in C^-(c_j^{i,[Hom]})$, $c_i^{ij,[HOB]} \in C^-(c_i^{ij,[Het]})$. For each $X \in$ {Hom, Het, HOB}, let $\hat{k}_i^{ij,[X]}$, $\hat{k}_j^{i,[X]}$ and $k_j^{i,[X]}$denote, respectively, the values taken by equations (3.4)-(3.5) when $c_j^i = c_j^{i,[X]}$ and $c_i^{ij} = c_i^{ij,[X]}$, and $k_i^{[X]}$ denote i's corresponding behavioral level. First compare treatment [Hom] and [Het]: since $c_i^{ij,[Het]} = c_i^{ij,[Hom]}$ and $c_j^{i,[Het]} \in C^-(c_j^{i,[Hom]})$, it follows that $\hat{k}_j^{i,[Het]} \leq \hat{k}_j^{i,[Hom]}$ and $\hat{k}_i^{ij,[Het]} \leq \hat{k}_i^{ij,[Hom]}$, which in turn implies $k_j^{i,[Het]} \leq k_j^{i,[Hom]}$ and hence $k_i^{[Het]} \leq k_i^{[Hom]}$; then compare [Het] and [HOB]: since $c_j^{i,[Het]} = c_j^{i,[HOB]}$, $\hat{k}_j^{i,[Het]} = \hat{k}_j^{i,[HOB]}$, but $c_i^{ij,[HOB]} \in C^-(c_i^{ij,[Het]})$ implies $\hat{k}_i^{ij,[HOB]} \leq \hat{k}_i^{ij,[Het]}$, which in turn implies $k_j^{i,[HOB]} \leq k_j^{i,[Het]}$ and hence $k_i^{[HOB]} \leq k_i^{[Het]}$ part (i) follows. The same argument also applies to $X+$, which implies part (iii).

(b) For parts (ii) and (iv), note that IA.3 implies, for any i with label II, we have $c_j^{i,[Hom]} \in C^-(c_j^{i,[Het]})$, $c_i^{ij,[Het]} \in C^-(c_i^{ij,[HOB]})$, and by IA.4 we have $c_j^{i,[Het]} \in C^+(c_i)$. Maintaining the same notation as above, first compare treatment [Hom] and [Het]: since $c_i^{ij,[Het]} = c_i^{ij,[Hom]}$ and $c_j^{i,[Het]} \in C^-(c_j^{i,[Hom]})$, it follows that $\hat{k}_j^{i,[Het]} \geq \hat{k}_j^{i,[Hom]}$ and $\hat{k}_i^{ij,[Het]} \geq \hat{k}_i^{ij,[Hom]}$, which in turn implies $k_j^{i,[Het]} \geq k_j^{i,[Hom]}$ and hence $k_i^{[Het]} \geq k_i^{[Hom]}$; under part (ii) of IA.4, we also have $c_j^{ij,[Het]} \in C^+(c_i)$, and hence $k_i^{[Het]} = \hat{k}_i^{[Het]}$. Next, compare [Het] and [HOB]: since $c_j^{i,[Het]} = c_j^{i,[HOB]}$, $\hat{k}_j^{i,[Het]} = \hat{k}_j^{i,[HOB]}$, but $c_i^{ij,[HOB]} \in C^-(c_i^{ij,[Het]})$ implies $\hat{k}_i^{ij,[Het]} \leq \hat{k}_i^{ij,[HOB]}$, which in turn implies $k_j^{i,[Het]} \leq k_j^{i,[HOB]}$ and hence $k_i^{[Het]} \leq k_i^{[HOB]}$. Since, by IA.3, $c_i^{ij,[HOB]} \in C^+\left(c_j^{i,[Het]}\right)$ and $c_j^{i,[Het]} \in C^+(c_i)$, we also have $k_i^{[HOB]} = \hat{k}_i^{[HOB]} = \hat{k}_i^{[Het]}$, and hence also $k_i^{[HOB]} = k_i^{[Het]}$ if part (ii) of IA.4 also holds. Part (ii) follows. The same argument also applies to $X+$, which implies part (iv).

2. IA.2 implies that $c_i^{ij,[AT-Het]} = c_i^{ij,[AT-Hom]}$ for any i of any label. IA.3 implies that for any i with label I, we have $c_j^{i,[AT-Het]} \in C^-(c_j^{i,[AT-Hom]})$. Maintaining the same notation as above, since $c_i^{ij,[AT-Het]} = c_i^{ij,[AT-Hom]}$ and $c_j^{i,[AT-Het]} \in C^-(c_j^{i,[AT-Hom]})$, it follows that $\hat{k}_j^{i,[AT-Het]} \leq \hat{k}_j^{i,[AT-Hom]}$ and $\hat{k}_i^{ij,[AT-Het]} \leq \hat{k}_i^{ij,[AT-Hom]}$, which in turn implies $k_j^{i,[AT-Het]} \leq k_j^{i,[AT-Hom]}$ and hence $k_i^{[AT-Het]} \leq k_i^{[AT-Hom]}$. It follows that $F_{AT-Het}^I \succsim F_{AT-Hom}^I$. The argument for $F_{AT-Hom}^{II} \succsim F_{AT-Het}^{II}$ is symmetric, swapping the inequalities, since IA.3 implies that for any i with label II, $c_j^{i,[AT-Hom]} \in C^-(c_j^{i,[AT-Het]})$, $c_i^{ij,[AT-Het]} \in C^-(c_i^{ij,[AT-HOB]})$.

3. The result follows from the same argument as in the previous point, just replacing AT-X with AP-X, for $X \in \{\text{Hom, Het}\}$.

□

Proof of Proposition 4. The result follows directly from the argument in the main text. □

3.7.2 Logistics of the Experiments

The experiments were conducted at the Laboratori d'Economia Experimental (LEEX) at Universitat Pompeu Fabra (UPF), Barcelona. Subjects were students of UPF, recruited using the LEEX system. No subject took part in more than one session. Subjects for the first experiment were paid 3 euros for showing up (students coming from a campus that was farther away received 4 euros instead). Subjects' earnings averaged 15.8. Subjects had a showup fee of 4 euros in the second experiment, and their earnings averaged 18 euros. The payments of subjects in Experiment 3 averaged at 14 euros.

Each subject in the first experiment went through a sequence of 18 games. Payoffs are expressed in 'tokens', each worth 15 cents. Subjects were paid randomly, once every six iterations. The order of treatments is randomized (see below). Subjects in the second and third experiment

each went through a sequence of 9 games, and were paid randomly based on three iterations. In those, to compensate for there being fewer games from which the payments were drawn, 8 tokens were worth 1 euro. For all experiments, subjects only observed their own overall earnings at the end, and received no information concerning their opponents' results.

Our subjects for the first experiment were divided in 6 sessions of 20 subjects, for a total of 120 subjects. Three sessions were based on the exogenous classification, and each contained 10 students from the field of humanities (humanities, human resources, and translation), and 10 from math and sciences (math, computer science, electrical engineering, biology and economics). Three sessions were based on the endogenous classification, and students were labeled based on their performance on a test of our design (see Alaoui and Penta (2016a)). In these sessions, half of the students were labeled as 'high' and half as 'low'.

There were 60 subjects for the second experiment and 34 for the third experiment. The subjects all took the endogenous classification test first but they were not given any feedback, and remained unlabeled.

3.7.2.1 Instructions of the Experiment

We describe in 3.7.2.1 the instructions as worded for a student from math and sciences in the first experiment. The instructions for students from humanities would be obtained replacing these labels everywhere. Similarly, labels high and low would be used for the endogenous classification. The related instructions for the second experiment are at the end of 3.7.2.1.

Baseline Game and Treatments [Hom], [Het] and [HOB]

Pick a number between 11 and 20. You will always receive the amount that you announce, in tokens.

In addition:

- If you give the same number as your opponent, you receive an extra 10 tokens.

- If you give a number that's exactly one less than your opponent, you receive an extra 20 tokens.

Example:

-If you say 17 and your opponent says 19, then you receive 17 and he receives 19.

-If you say 12 and your opponent says 13, then you receive 32 and he receives 13.

-If you say 16 and your opponent says 16, then you receive 26 and he receives 26.

Treatments [Hom] and [Het]:

Your opponent is:

- a student from maths and sciences (treatment [Hom]) / humanities (treatment [Het])

- he is given the same rules as you.

Treatment [HOB]:

In this case, the number you play against is chosen by:

- a student from humanities facing another student from humanities. In other words, two students from humanities play against each other. You play against the number that one of them has picked.

Treatment [Hom-Rep]

In this case, the number you play against is chosen by:

- a student from maths and sciences facing a student from maths and sciences. In other words, two students from maths and sciences play against each other. You play against the number that one of them has picked.

Changing Payoffs: Treatments [HOM+], [Het+], [HOB+] and [Hom-Rep]

You are now playing a high-payoff game. Pick a number between 11 and 20. You will always receive the amount that you announce, in tokens.

In addition:

- If you give the same number as your opponent, you receive an extra 10 tokens.

- If you give a number that's exactly one less than your opponent, you receive an extra 80 tokens.

Example:

-If you say 17 and your opponent says 19, then you receive 17 and he receives 19.

-If you say 12 and your opponent says 13, then you receive 92 and he receives 13.

-If you say 16 and your opponent says 16, then you receive 26 and he receives 26.

Treatments [Hom+] and [Het+]

Your opponent is:

- a student from maths and sciences playing the high-payoff game (treatment [Hom+]) / humanities (treatment [Het+])

- he is given the same rules as you.

Treatment [HOB+]

In this case, the number you play against is chosen by:

- a student from humanities playing the high payoff game with another student from humanities. In other words, two students from humanities play the high payoff game with each other (extra 10 if they tie, 80 if exactly one less than opponent). You play against the number that one of them has picked.

Treatments [Tut], [AT-Hom] and [AT-Het]

Before playing treatments [Tut], [AT-Hom] and [AT-Het], the subjects were given the 'tutorial' stated in Section 3.3.1.2.

Treatment [Tut]

Your opponent is:

- a student who has also been given the game theory tutorial.

Treatment [AT-Hom]

Your opponent is:

- a student from maths and sciences,

- he has not been given the game theory tutorial.

Treatment [AT-Het]

Your opponent is:

- a student from humanities,

- he has not been given the game theory tutorial.

Treatments [AP-Hom] and [AP-Het]

Treatment [AP-Hom] You are now playing the high-payoff game.

In this case, the number you play against is chosen by:

- a student from maths and sciences playing the low payoff game with another student from maths and sciences. In other words, two students from maths and sciences play the low payoff game with each other (extra 10 if they tie, 20 if exactly one less than opponent). You play against the number that one of them has picked.

Treatment [AP-Het] You are now playing the high-payoff game.

In this case, the number you play against is chosen by:

- a student from humanities playing the low payoff game with another student from humanities. In other words, two students from humanities play the low payoff game with each other (extra 10 if they tie, 20 if exactly one less than opponent). You play against the number that one of them has picked.

Treatments [Un], [Un+], [AP-Un], [Tut-Un], [AT-Un]

The treatments for the second experiment contain no information concerning own or opponents' label, and are adjusted accordingly. The third experiment contains identical treatments.

Treatments **[Un]**, **[Un+]** and **[AP-Un]** are identical to **[Hom]** (and **Het**), **[Hom+]** (and **Het+**) and **[AP-Hom]** (and **AP-Het**), respectively, of the first experiment, but with the following information concerning the opponent:

Your opponent is given the same rules as you.

Treatment **[Tut-Un]** is also preceded by the same game theory tutorial as **[Tut]** and the same game, followed by:

Your opponent has also seen the game theory tutorial.

Treatment **[AT-Un]** is identical to treatment **[AT-Hom]** (and **AT-Het**), with the following information concerning the opponent:

In this case, you are playing against a subject who has not seen the game theory tutorial, and who himself (or herself) plays against a subject who hasn't seen the tutorial either. In other words, the two subjects have played one another. You play against the number that one of them has chosen.

3.7.2.2 Sequences

In the first experiment, our 6 groups (3 for the endogenous and 3 for the exogenous classification) went through four different sequences of treatments. Two of the groups in the exogenous treatment followed Sequence 1, and one followed Sequence 2. The three groups of the endogenous classification each took a different sequence: respectively sequence 1, 3 and 4. All the sequences contain the treatments [Hom], [Het], [HOB] ,[Hom+], [Het+], [HOB+]. The order of the main treatments is different in each sequence, both in terms of changing the beliefs and the payoffs. Some sequences include treatments [AP-Hom], [AP-Het] while others included [Tut], [AT-Hom] and [AT-Het].

- **Sequence 1:** Hom, Het, HOB, Het, Hom, HOB, Hom+, Het+, HOB+, Het+, Hom+, HOB+, Tut, AT-Hom, AT-Het, Tut, AT-Hom, AT-Het

- **Sequence 2:** Hom, Het, Het, Hom, HOB, HOB, AP-Hom, AP-Hom, AP-Hom, AP-Hom, AP-Hom, AP-Hom, Hom+, Het+, Het+, Hom+, HOB+, HOB+

- **Sequence 3:** Hom+, Het+, HOB+, Het+, Hom+, HOB+, Hom, Het, HOB, Het, Hom, HOB, Tut, AT-Hom, AT-Het, Tut, AT-Hom, AT-Het

- **Sequence 4:** Het, Hom, HOB, Hom, Het, HOB, AP-Hom, AP-Het, AP-Hom, AP-Het, AP-Hom, AP-Het, Het+, Hom+, HOB+, Hom+, Het+, HOB+

- **Sequence Robustness:** Hom, Het, HOB, Hom-Rep, Het, Hom, HOB, Hom-Rep, AP-Hom, AP-Het, AP-Hom, AP-Het, Het+, Hom+, HOB+, Hom+, Het+, HOB+

The second and third experiments contained unlabeled treatments only and subjects went through the following sequence:

- **Sequence Unlabeled:** Un, Un+, AP-Un, AP-Un, Un+, Tut-Un, AT-Un, Tut-Un, AT-Un

3.7.2.3 Details of the Cognitive Test

The cognitive test that subjects played in Experiment 1 takes roughly 30 minutes to complete, and consists of three questions. First, subjects are asked to play a computerised version of the board game Mastermind. Second, subjects are given a typical centipede game of seven rounds, and are

asked what an infinitely sophisticated and rational agent would do. Third, subjects are given a lesser known 'pirates game', which is a four player game that can be solved by backward induction. Subjects are asked what the outcome of this game would be, if players were 'infinitely sophisticated and rational'. Each question was given a score, and then a weighted average was taken. Subjects whose score was higher (lower) than the median score were labeled as 'high' ('low'). Subjects in Experiments 2 and 3 saw an additional question before the other questions. This question was a 'muddy faces' game where subjects had to perform iterated reasoning to answer the three sub-questions correctly. We report next the instructions of the test, as administered to the students (see the online appendix for the original version in Spanish).

Instructions of the test. This test consists of three questions. You must answer all three within the time limit stated.

Question 1:

In this question, you have to guess four numbers in the correct order. Each number is between 1 and 7. No two numbers are the same. You have nine attempts to guess the four numbers. After each attempt, you will be told the number of correct answers in the correct place, and the number of correct numbers in the wrong place.

Example: Suppose that the correct number is: 1 4 6 2.

If you guess: 3 5 4 6, then you will be told that you have 0 correct answers in the correct place and 2 in the wrong place.

If you guess: 3 5 6 4, then you will be told that you have 1 correct answer in the correct place and 1 in the wrong place.

If you guess: 3 4 7 2, then you will be told that you have 2 correct answers in the correct place and 0 in the wrong place.

If you guess: 1 4 6 2, then you will be told that you have 4 correct answers, and you have reached the objective.

Notice that the correct number could not be (for instance) 1 4 4 2, as 4 is repeated twice. You are, however, allowed to guess 1 4 4 2, in any round.

You have a total of 90 seconds per round: 30 seconds to introduce the numbers and 60 seconds to view the results.

Question 2:

Consider the following game. Two people, Antonio and Beatriz, are moving sequentially. The game starts with 1 euro on the table. There are at most 6 rounds in this game:

Round 1) Antonio is given the choice whether to take this 1 euro, or pass, in which case the game has another round. If he takes the euro, the game ends. He gets 1 euro, Beatriz gets 0 euros. If Antonio passes, they move to round 2.

Round 2) 1 more euro is put on the table. Beatriz now decides whether to take 2 euros, or pass. If she takes the 2 euros, the game ends. She receives 2 euros, and Antonio receives 0 euros. If Beatriz passes, they move to round 3.

Round 3) 1 more euro is put on the table. Antonio is asked again: he can either take 3 euros and leave 0 to Beatriz, or pass. If Antonio passes, they move to round 4.

Round 4) 1 more euro is put on the table. Beatriz can either take 3 euros and leave 1 euro to Antonio, or pass. If Beatriz passes, they move to round 5.

Round 5) 1 more euro is put on the table. Antonio can either take 3 euros and leave 2 to Beatriz, or pass. If Antonio passes, they move to round 6.

Round 6) Beatriz can either take 4 euros and leaves 2 to Antonio, or she passes, and they both get 3.

Assume Antonio and Beatriz are infinitely sophisticated and rational and they each want to get as much money as possible. What will be the outcome of the game?

a) Game stops at Round 1, with payoffs:	(Antonio: 1 euro	Beatriz: 0 euros)
b) Game stops at Round 2, with payoffs:	(Antonio: 0 euro	Beatriz: 2 euros)
c) Game stops at Round 3, with payoffs:	(Antonio: 2 euros	Beatriz: 1 euro)
d) Game stops at Round 4, with payoffs:	(Antonio: 1 euro	Beatriz: 3 euros)
e) Game stops at Round 5, with payoffs:	(Antonio: 3 euros	Beatriz: 2 euros)
f) Game stops at Round 6, with payoffs:	(Antonio: 2 euros	Beatriz: 4 euros)
g) Game stops at Round 6, with payoffs:	(Antonio: 3 euros	Beatriz: 3 euros)

You have 8 minutes in total for this question.

Question 3:

Four pirates (Antonio, Beatriz, Carla and David) have obtained 10 gold doblones and have to divide up the loot. Antonio proposes a distribution of the loot. All pirates vote on the proposal. If half the crew or more agree, the loot is divided as proposed by Antonio.

If Antonio fails to obtain support of at least half his crew (including himself), then he will be killed. The pirates start over again with Beatriz as the proposer. If she gets half the crew (including herself) to agree, then the loot is divided as proposed. If not, then she is killed, and Carla then makes the proposal. Finally, if her proposal is not agreed on by half the people left, including herself, then she is killed, and David takes everything.

In other words:

Antonio needs 2 people (including himself) to agree on his proposal, and if not he is killed.

If Antonio is killed, Beatriz needs 2 people (including herself) to agree on her proposal, if not she is killed.

If Beatriz is killed, Carla needs 1 person to agree (including herself) to agree on her proposal, and if not she is killed.

If Carla is killed, David takes everything.

The pirates are infinitely sophisticated and rational, and they each want to get as much money as possible. What is the maximum number of coins Antonio can keep without being killed?

Notice that *the proposer* can also vote, and that exactly half the votes is enough for the proposal to pass.

You have 8 minutes in total for this question.

Scoring in Experiment 1. In the mastermind question, subjects were given 100 points if correct, otherwise they received 15 points for each correct answer in the correct place and 5 for each correct answer in the wrong place in their last answer. In the centipede game, subjects were given 100 points if they answered that the game would end at round 1, otherwise points were equal to $\min\{0, (6 - \text{round}) \times 15\}$. In the pirates game, subjects obtain 100 if they answer 9, 60 if they answer 10, and $\max\{0, (x - 2) * 10\}$ otherwise. The overall score was given by the average of the three.

Question 0 (Experiments 2 and 3):

There are three people, A, B and C, each with a circle on their forehead. The circle can be white or black. Every person can see the circle on the others' forehead but not the one on their own. In reality, A and C have a white circle and B has a black circle:

They are given the following instructions, in this order, and can observe the reaction of the others:

If you know that your circle is black, take a step forward. Who will take a step forward?

Now, they are informed that at least one of them has a black circle. They are then asked: If you know the color of your circle, take a step forward. Who will take a step forward?

They observe the reaction to the previous question (in other words, they see who took a step forward). They are asked: Now that you have seen who stepped forward, if you know the color of your circle, take a step forward. Who will take a step forward? (Include only those new persons who take a step forward, don't include anyone who already took a step forward in the previous questions.)

Scoring in Experiments 2 and 3. Scoring for the three common games was the same as in Experiment 1. The muddy faces game gave a total of 120 points if the correct answer was given in each sub-question. Partial points were given depending on how close the answer was to the correct iterative reasoning. In order to calculate the overall score, the scores for each of these sub-questions were added to those from the three other questions and the resulting sum was divided by 4.2.

3.7.3 Regressions

For all following tables, the number of observations refers to the number of treatment observations. These observations are clustered at the id level into groups. For example, for Table 3.10, there are 235 treatment observations clustered at the id level to form 59 groups (i.e. 59 subjects played these treatments). All standard errors are thus clustered at the individual level, taking into account the dependence of the treatment outcomes.
The general regression equation used for these panel regressions is the

following:

$$Y_{i,t,k,l} = \alpha_{k,l} + \beta_{k,l} Treat_{i,t,k,l} + \gamma_{k,l} Endog_{i,k,l} + \epsilon_{i,t,k,l}$$

where i denotes the individuals, t denotes the period in which the treatment was played, k denotes the treatments of interest and l the label of interest (this subscript does not appear in the regression equations of experiments 2 and 3). For example, for the first regression in Table 3.10, $k = [Hom]$or$[Hom+]$ i.e. all treatments that are neither [Hom] nor [Hom+] are dropped from this regression. All choices that were made by each individual in each period in treatments [Hom] and [Hom+] are captured by $Y_{i,t,k,l}$, while $Treat_{i,t,k,l}$ is a dummy that takes value 1 if the treatment is [Hom+] and 0 if the treatment was [Hom]. $l = LabelI$ and subjects from the other label are dropped from the regression. We include time effects as treatments were repeated. For Experiment 1, the regressions also include a "classification dummy" that takes value 1 if the subject was in the endogenous classification group and 0 if they were the exogenous classification group. This dummy is not time varying. For Table 3.13, instead of splitting the sample by labels, a label I dummy is included as well as an interaction term between the treatment and the label dummies.

Table 3.10: Experiment 1, Regressions on Payoffs Effects (joint for all sequences)

VARIABLES	Relevant Dummy	Classification Dummy	Constant	Observations
From [Hom] to [Hom+], Label I	-0.50***	0.22	17.21***	235
	(0.19)	(0.52)	(0.33)	
From [Hom] to [Hom+], Label II	-0.64**	-0.10	16.92***	236
	(0.32)	(0.46)	(0.34)	
From [Het] to [Het+], Label I	-0.62***	0.37	17.50***	233
	(0.21)	(0.38)	(0.33)	
From [Het] to [Het+], Label II	-0.74***	0.39	16.58***	236
	(0.28)	(0.48)	(0.39)	
From [HOB] to [HOB+], Label I	-1.15***	0.34	17.77***	236
	(0.24)	(0.38)	(0.30)	
From [HOB] to [HOB+], Label II	-0.97***	-0.07	16.97***	234
	(0.26)	(0.45)	(0.37)	

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

Table 3.11: Experiment 1, Regressions from Post-tutorial treatments.

VARIABLES	Relevant Dummy	Classification Dummy	Constant	Observations
From [Hom] to [AT-Hom], Label *I*	-0.71	0.47	16.67***	156
	(0.47)	(0.67)	(0.49)	
From [Hom] to [AT-Hom], Label *II*	-0.85**	0.32	17.09***	156
	(0.42)	(0.66)	(0.46)	
From [Het] to [AT-Het], Label *I*	-0.32	0.58	17.29***	156
	(0.31)	(0.54)	(0.48)	
From [Het] to [AT-Het], Label *II*	-1.92***	0.30	17.04***	156
	(0.49)	(0.72)	(0.54)	
From [AT-Hom] to [AT-Het], Label *I*	1.06***	0.92	15.73***	156
	(0.36)	(0.65)	(0.57)	
From [AT-Hom] to [AT-Het], Label *II*	-1.14***	-0.16	16.49***	156
	(0.40)	(0.88)	(0.56)	

Clustered standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 3.12: Experiment 1, Regressions for asymmetric payoff treatments.

VARIABLES	Relevant Dummy	Classification Dummy	Constant	Observations
From [Hom] to [AP-Hom], Label *I*	-0.74***	0.88*	17.69***	100
	(0.11)	(0.47)	(0.42)	
From [AP-Hom] to [Hom+], Label *I*	-0.11	0.88	16.94***	100
	(0.17)	(0.45)	(0.35)	
From [Hom] to [AP-Hom], Label *II*	-0.38	-0.18	16.19***	100
	(0.55)	(0.69)	(0.50)	
From [AP-Hom] to [Hom+], Label *II*	-0.24	-0.52	15.98***	100
	(0.37)	(0.81)	(0.48)	
From [HOB] to [AP-Het], Label *I*	-0.50***	0.68	17.86***	99
	(0.14)	(0.46)	(0.37)	
From [AP-Het] to [HOB+], Label *I*	-0.55*	0.60	17.40***	99
	(0.29)	(0.41)	(0.39)	
From [HOB] to [AP-Het], Label *II*	-0.28	-0.28	16.49***	100
	(0.36)	(0.67)	(0.54)	
From [AP-Het] to [HOB+], Label *II*	-0.32	-0.52	16.33***	100
	(0.25)	(0.73)	(0.47)	
From [AP-Hom] to [AP-Het], Label *I*	0.31**	0.64	17.06***	119
	(0.13)	(0.45)	(0.38)	
From [AP-Hom] to [AP-Het], Label *II*	0.35	-0.28	15.86***	120
	(0.24)	(0.72)	(0.48)	

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

Table 3.13: Experiment 1, regressions to examine effect of labels for going "From treatment x to y" (using a dummy for Label I, treatment dummy, and a label-treatment interaction term).

VARIABLES	Label I dummy	Treatment dummy 0: if treatment x 1: if treatment y	Interaction	Constant	Obs.
From [Hom] to [AP-Hom]	2.03***	-0.38	-0.36	16.10***	200
	(0.57)	(0.54)	(0.55)	(0.49)	
From [AP-Hom] to [Hom+]	1.67***	-0.24	0.13	15.72***	200
	(0.45)	(0.36)	(0.40)	(0.38)	
From [HOB] to [AP-Het]	1.85***	-0.28	-0.22	16.35***	199
	(0.45)	(0.35)	(0.38)	(0.38)	
From [AP-Het] to [HOB+]	1.63***	-0.32	-0.23	16.07***	199
	(0.45)	(0.24)	(0.38)	(0.37)	
From [AP-Hom] to [AP-Het]	1.67***	0.35	-0.04	15.72***	239
	(0.45)	(0.24)	(0.27)	(0.38)	
From [Hom] to [AT-Hom]	-0.35	-0.85**	0.14	17.26***	312
	(0.50)	(0.42)	(0.61)	(0.31)	
From [Het] to [AT-Het]	0.40	-1.92***	1.60***	17.19***	312
	(0.44)	(0.49)	(0.57)	(0.34)	
From [AT-Hom] to [AT-Het]	-0.21	-1.14***	2.21***	16.41***	312
	(0.61)	(0.40)	(0.53)	(0.45)	

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

Table 3.14: Experiment 2, regressions for all treatments.

VARIABLES	Relevant dummy	Constant	Observations
From [Un] to [AT-Un]	-0.22	17.66***	173
	(0.32)	(0.28)	
From [Un] to [Un+]	-1.10***	17.68***	174
	(0.32)	(0.28)	
From [Un] to [AP-Un]	-0.52*	17.66***	172
	(0.28)	(0.28)	
From [AP-Un] to [Un+]	-0.55**	17.13***	232
	(0.28)	(0.25)	

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

Table 3.15: Experiment 3, regressions for all treatments.

VARIABLES	Relevant dummy	Constant	Observations
From [E3-Un] to [E3-AT-Un]	-1.03*	17.75***	90
	(0.60)	(0.42)	
From [E3-Un] to [E3-Un+]	-1.01***	17.72***	87
	(0.37)	(0.41)	
From [E3-Un] to [E3-AP-Un]	-1.02**	17.70***	83
	(0.41)	(0.42)	
From [E3-AP-Un] to [E3-Un+]	-0.004	16.72***	112
	(0.40)	(0.38)	

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

Table 3.16: Experiments 1-3, Wilcoxon signed rank test results for all treatments.

Experiment	Treatment	P-value
1	[Hom] to [Hom+], Label I	0.009***
	[Hom] to [Hom+], Label II	0.036**
	[Het] to [Het+], Label I	0.001***
	[Het] to [Het+], Label II	0.007***
	[HOB] to [HOB+], Label I	0.000***
	[HOB] to [HOB+], Label II	0.001***
	[Hom] to [AT-Hom], Label I	0.074*
	[Hom] to [AT-Hom], Label II	0.052*
	[Het] to [AT-Het], Label I	0.269
	[Het] to [AT-Het], Label II	0.000***
	[AT-Hom] to [AT-Het], Label I	0.015**
	[AT-Hom] to [AT-Het], Label II	0.010***
	[Hom] to [AP-Hom], Label I	0.000***
	[Hom] to [AP-Hom], Label II	0.331
	[AP-Hom] to [Hom+], Label I	0.601
	[AP-Hom] to [Hom+], Label II	0.435
	[HOB] to [AP-Het], Label I	0.002***
	[HOB] to [AP-Het], Label II	0.438
	[AP-Het] to [HOB+], Label I	0.021**
	[AP-Het] to [HOB+], Label II	0.055*
	[AP-Hom] to [AP-Het], Label I	0.029**
	[AP-Hom] to [AP-Het], Label II	0.211
2	[Un] to [AT-Un]	0.033**
	[Un] to [Un+]	0.000***
	[Un] to [AP-Un]	0.003***
	[AP-Un] to [Un+]	0.003***
3	[E3-Un] to [E3-AT-Un]	0.042**
	[E3-Un] to [E3-Un+]	0.016**
	[E3-Un] to [E3-AP-Un]	0.000***
	[E3-AP-Un] to [E3-Un+]	0.900

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

3.7.4 Additional Figures

3.7.4.1 TOM scores

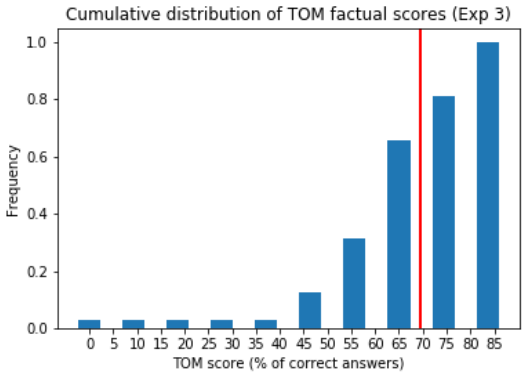


Figure 3.11: Distribution of factual TOM scores. Vertical red line indicates sample average.

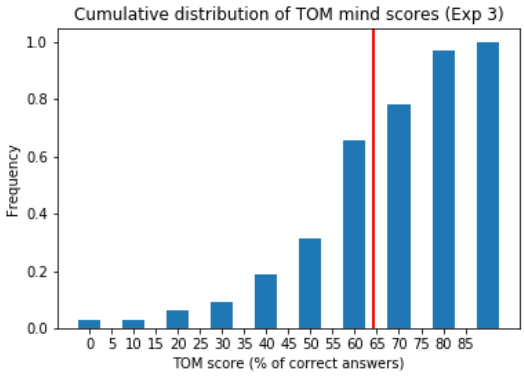


Figure 3.12: Distribution of TOM mind scores. Vertical red line indicates sample average.

3.7.4.2 Individual behavior: violations of theory

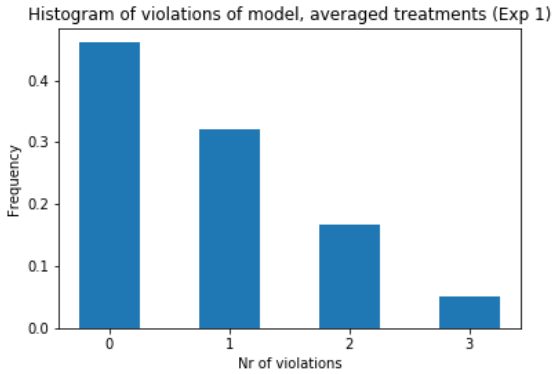


Figure 3.13: Experiment 1 (tutorial sessions): Number of violations of theory for averaged treatments.

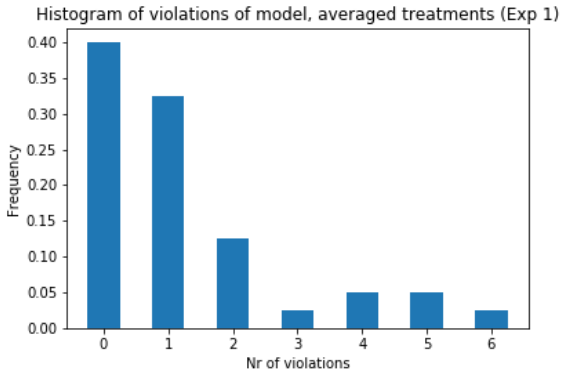


Figure 3.14: Experiment 1 (payoff sessions): Number of violations of theory for averaged treatments.

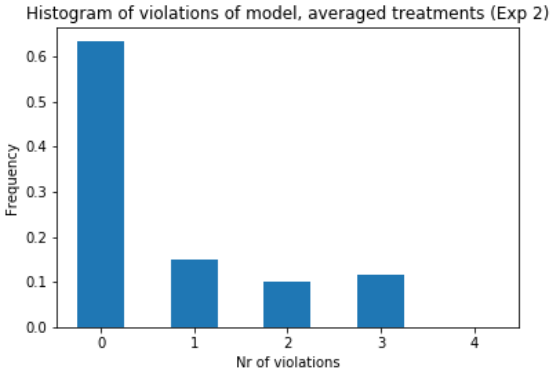


Figure 3.15: Experiment 2: Number of violations of theory for averaged treatments.

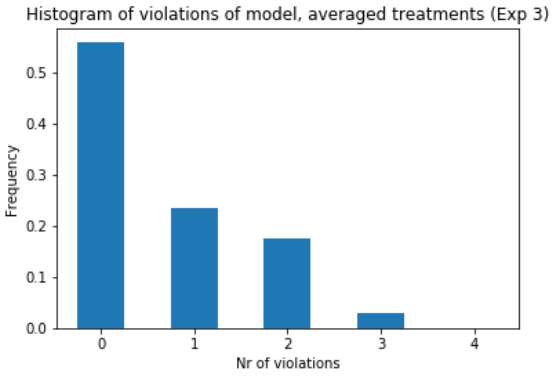


Figure 3.16: Experiment 3: Number of violations of theory for averaged treatments.

3.7.4.3 Individual behavior: shifts in behavior

Experiment 1: Label I, HOB to AP-Het

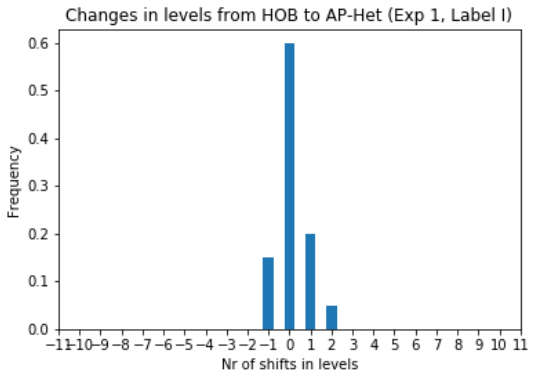


Figure 3.17: Experiment 1: Frequency of shifts in level played from [HOB] to [AP-Het].

Experiment 3

(a) (b)

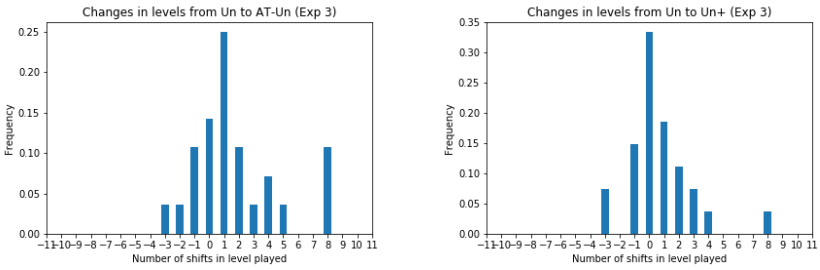


Figure 3.18: Experiment 3: Frequency of shifts in level played from [Un] to [AT-Un] (left) and from [Un] to [Un+] (right).

(a) (b)

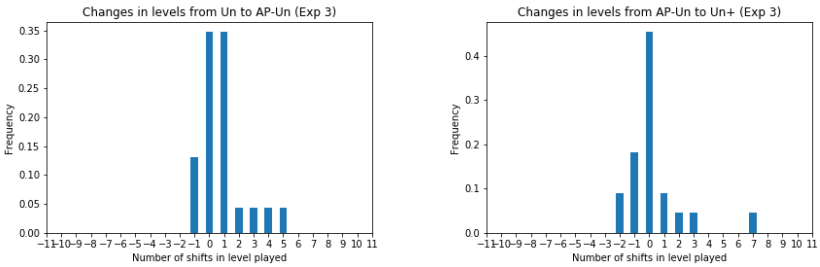


Figure 3.19: Experiment 3: Frequency of shifts in level played from [Un] to [AP-Un] (left) and from [AP-Un] to [Un+] (right).

Bibliography

Abeler, J., A. Becker, and A. Falk (2014): "Representative evidence on lying costs," *Journal of Public Economics*, 113, 96–104.

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for truth-telling," *Econometrica*, 87, 1115–1153.

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman. (2012): "Beliefs and Endogenous Cognitive Levels: An Experimental Study," *Games and Economic Behavior*, 75, 449–63.

Alaoui, L., K. A. Janezic, and A. Penta (2020): "Reasoning about others' reasoning," *Journal of Economic Theory*, 189, 105091.

Alaoui, L. and A. Penta (2012): "Level-k Reasoning and Incentives," *mimeo*.

——— (2016a): "Endogenous Depth of Reasoning," *Review of Economic Studies*, 83, 1297–1333.

——— (2016b): "Endogenous Depth of Reasoning and Response Time, with an application to the Attention-Allocation Task," *mimeo*.

——— (2018): "Cost-Benefit Analysis in Reasoning," *Barcelona GSE working paper n.1062*.

Arad, A. and A. Rubinstein (2012): "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review*, 102, 3561–3573.

Avoyan, A. and A. Schotter (2020): "Attention in games: An experimental study," *European Economic Review*, 124, 103410.

Bagnall, A., G. Janacek, and M. Zhang (2003): "Clustering time series from mixture polynomial models with discretised data," .

Basu, K. (1991): "The Traveler's Dilemma: Paradoxes of Rationality in Game Theory," *American Economic Review Papers and Proceedings*, 84, 391–395.

Besley, T. (2005): "Political selection," *Journal of Economic Perspectives*, 19, 43–60.

Besley, T., O. Folke, T. Persson, and J. Rickne (2017): "Gender quotas and the crisis of the mediocre man: Theory and evidence from Sweden," *American Economic Review*, 107, 2204–42.

Biziou-van Pol, L., J. Haenen, A. Novaro, A. Occhipinti Liberman, and V. Capraro (2015): "Does telling white lies signal pro-social preferences?" *Judgment and Decision Making*, 10, 538–548.

Bovet, A. and H. A. Makse (2019): "Influence of fake news in Twitter during the 2016 US presidential election," *Nature Communications*, 10, 7.

Camerer, C. F. (2003): *Behavioral Game Theory*, Princeton University Press, Princeton, NJ, USA.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 119, 861–898.

Cameron, S. V. and J. J. Heckman (1998): "Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males," *Journal of Political economy*, 106, 262–333.

Cappelen, A. W., E. Ø. Sørensen, and B. Tungodden (2013): "When do we lie?" *Journal of Economic Behavior & Organization*, 93, 258–265.

Capra, C. M., J. K. Goeree, R. Gomez, and C. A. Holt. (1999): "Anomalous Behavior in a Traveler's Dilemma?" *American Economic Review*, 89, 678–690.

Capraro, V. (2017): "Does the truth come naturally? Time pressure increases honesty in one-shot deception games," *Economics Letters*, 158, 54–57.

——— (2018): "Gender differences in lying in sender-receiver games: A meta-analysis," *Judgment and Decision Making*, 13, 345–355.

Capraro, V., J. Schulz, and D. G. Rand (2019): "Time pressure and honesty in a deception game," *Journal of Behavioral and Experimental Economics*, 79, 93–99.

Caselli, F. and M. Morelli (2004): "Bad politicians," *Journal of Public Economics*, 88, 759–782.

Chamberlain, G. (1983): "Funds, factors, and diversification in arbitrage pricing models," *Econometrica*, 1305–1323.

Charron, N., L. Dijkstra, and V. Lapuente (2014): "Regional governance matters: Quality of government within European Union member states," *Regional Studies*, 48, 68–90.

Cohn, A., E. Fehr, and M. A. Maréchal (2014): "Business culture and dishonesty in the banking industry," *Nature*, 516, 86–89.

Cohn, A. and M. A. Maréchal (2018): "Laboratory measure of cheating predicts school misconduct," *The Economic Journal*, 128, 2743–2754.

Costa-Gomes, M. A. and V. P. Crawford (2006): "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, 96, 1737–1768.

Costa-Gomes, M. A., V. P. Crawford, and B. Broseta. (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69, 1193–1235.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013): "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, 51.

Dai, Z., F. Galeotti, and M. C. Villeval (2017): "Cheating in the lab predicts fraud in the field: An experiment in public transportation," *Management Science*, 64, 1081–1100.

Dal Bó, E., F. Finan, O. Folke, T. Persson, and J. Rickne (2017): "Who becomes a politician?" *The Quarterly Journal of Economics*, 132, 1877–1914.

De Vries, C. E. and H. Solaz (2017): "The electoral consequences of corruption," *Annual Review of Political Science*, 20, 391–408.

della Porta, D. (2004): "Political parties and corruption: Ten hypotheses on five vicious circles," *Crime, Law and Social Change*, 42, 35–60.

Dreber, A. and M. Johannesson (2008): "Gender differences in deception," *Economics Letters*, 99, 197–199.

Engle, R. and M. Watson (1981): "A one-factor multivariate time series model of metropolitan wage rates," *Journal of the American Statistical Association*, 76, 774–781.

Erat, S. and U. Gneezy (2012): "White lies," *Management Science*, 58, 723–733.

Falat, L. and L. Pancikova (2015): "Quantitative modelling in economics with advanced artificial neural networks," *Procedia economics and finance*, 34, 194–201.

Fearon, J. D. (1999): "Electoral accountability and the control of politicians: Selecting good types versus sanctioning poor performance," in *Democracy, Accountability, and Representation*, ed. by A. Przeworski, S. Stokes, and B. Manin, Cambridge: Cambridge University Press, 55–97.

Ferraz, C. and F. Finan (2008): "Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes," *The Quarterly Journal of Economics*, 123, 703–745.

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in disguise: an experimental study on cheating," *Journal of the European Economic Association*, 11, 525–547.

Focardi, S. M. and F. J. Fabozzi (2001): "Clustering economic and financial time series: Exploring the existence of stable correlation conditions," *Discussion Paper*.

Frederick, S. (2005): "Cognitive reflection and decision making," *Journal of Economic perspectives*, 19, 25–42.

Friedenberg, A., W. Kets, and T. Kneeland (2017): "Bounded Reasoning: Rationality or Cognition," *mimeo*.

Gächter, S. and J. F. Schulz (2016): "Intrinsic honesty and the prevalence of rule violations across societies," *Nature*, 531, 496–499.

Georganas, S., P. J. Healy, and R. A. Weber (2015): "On the persistence of strategic sophistication," *Journal of Economic Theory*, 159, 369–400.

Gibson, R., C. Tanner, and A. F. Wagner (2013): "Preferences for truthfulness: Heterogeneity among and within individuals," *The American Economic Review*, 103, 532–548.

Gill, D. and V. Prowse (2016): "Cognitive ability and learning to play equilibrium: A level-k analysis," *Journal of Political Economy*, 126, 1619–1676.

——— (2017): "Strategic complexity and the value of thinking," *mimeo*.

Gneezy, U. (2005): "Deception: The role of consequences," *The American Economic Review*, 95, 384–394.

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013): "Measuring lying aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

Goeree, J. K. and C. A. Holt (2001): "Ten Little Treasures of Game Theory and Ten Intuitive Contradictions," *American Economic Review*, 91, 1402–1422.

Goeree, J. K. and C. A. Holt. (2004): "A Model of Noisy Introspection," *Games and Economic Behavior*, 46, 365–382.

Goeree, J. K., P. Louis, and J. Zhang (2017): "Noisy Introspection in the 11-20 Game," *Economic Journal*.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer (2019): "Fake news on Twitter during the 2016 US presidential election," *Science*, 363, 374–378.

Hafner-Burton, E. M., D. A. Hughes, and D. G. Victor (2013): "The cognitive revolution and the political psychology of elite decision making," *Perspectives on Politics*, 11, 368–386.

Hanna, R. and S.-Y. Wang (2017): "Dishonesty and selection into public service: Evidence from India," *American Economic Journal: Economic Policy*, 9, 262–90.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Heifetz, A. and W. Kets (2018): "Robust multiplicity with a grain of naiveté," *Theoretical Economics*, 13, 415–465.

Heywood, P. M. (2007): "Corruption in contemporary Spain," *PS: Political Science & Politics*, 40, 695–699.

Hurkens, S. and N. Kartik (2009): "Would I lie to you? On social preferences and lying aversion," *Experimental Economics*, 12, 180–192.

Jacobsen, C., T. R. Fosgaard, and D. Pascual-Ezama (2018): "Why do we lie? A practical guide to the dishonesty literature," *Journal of Economic Surveys*, 32, 357–387.

Janezic, K. A. and A. Gallego (2020): "Eliciting preferences for truth-telling in a survey of politicians," *Proceedings of the National Academy of Sciences*, 117, 22002–22008.

——— (Deposited June 7 2020): "Replication Data for: "Eliciting preferences for truth-telling in a survey of politicians"," *Harvard Dataverse*, https://doi.org/10.7910/DVN/MPAZUD.

Joly, J., S. Soroka, and P. Loewen (2018): "Nice guys finish last: Personality and political success," *Acta Politica*, 54, 667–683.

Kajackaite, A. and U. Gneezy (2017): "Incentives and cheating," *Games and Economic Behavior*, 102, 433–444.

Kartik, N. (2009): "Strategic communication with lying costs," *The Review of Economic Studies*, 76, 1359–1395.

Kerschbamer, R., D. Neururer, and A. Gruber (2019): "Do altruists lie less?" *Journal of Economic Behavior & Organization*, 157, 560–579.

Kets, W. (2017): "Bounded Reasoning and Higher-Order Uncertainty," *mimeo*.

Kneeland, T. (2015): "Identifying Higher-order Rationality," *Econometrica*, 83, 2065–2079.

Levine, E. E. and M. E. Schweitzer (2014): "Are liars ethical? On the tension between benevolence and honesty," *Journal of Experimental Social Psychology*, 53, 107–117.

Liddle, B. and D. Nettle (2006): "Higher-order Theory of Mind and Social Competence in School-age Children," *Journal of Cultural and Evolutionary Psychology*, 4, 231–246.

Lohse, T., S. A. Simon, and K. A. Konrad (2018): "Deception under time pressure: Conscious decision or a problem of awareness?" *Journal of Economic Behavior & Organization*, 146, 31–42.

Maravall, J. M. (1999): "Accountability and Manipulation," in *Democracy, Accountability, and Representation*, ed. by A. Przeworski, S. Stokes, and B. Manin, Cambridge: Cambridge University Press, 154–196.

McLeod, B. A. and R. L. Genereux (2008): "Predicting the acceptability and likelihood of lying: The interaction of personality with type of lie," *Personality and Individual Differences*, 45, 591–596.

Mondak, J. J. (1995): "Competence, integrity, and the electoral success of congressional incumbents," *The Journal of Politics*, 57, 1043–1069.

Nagel, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 1313–1326.

Palan, S. and C. Schitter (2018): "Prolific. ac?A subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, 17, 22–27.

Peer, E., L. Brandimarte, S. Samat, and A. Acquisti (2017): "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, 70, 153–163.

Potters, J. and J. Stoop (2016): "Do cheaters in the lab also cheat in the field?" *European Economic Review*, 87, 26–33.

Proto, E., A. Rustichini, and A. Sofianos (2019): "Intelligence, Personality and Gains from Cooperation in Repeated Interactions," *Journal of Political Economy*, 127, 1351–1390.

Pujas, V. and M. Rhodes (1999): "Party finance and political scandal in Italy, Spain and France," *West European Politics*, 22, 41–63.

Rammstedt, B. and O. P. John (2007): "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of research in Personality*, 41, 203–212.

Rampal, J. (2018a): "Limited Foresight Equilibrium," *mimeo*.

——— (2018b): "Opponent's Foresight and Optimal Choice," *mimeo*.

Rosenbaum, S. M., S. Billinger, and N. Stieglitz (2014): "Let's be honest: A review of experimental evidence of honesty and truth-telling," *Journal of Economic Psychology*, 45, 181–196.

Sheffer, L., P. J. Loewen, S. Soroka, S. Walgrave, and T. Sheafer (2018): "Nonrepresentative representatives: An experimental study of the decision making of elected politicians," *American Political Science Review*, 112, 302–321.

Stahl, D. and P. Wison (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218–254.

Stiller, J. and R. I. M. Dunbar (2007): "Perspective-taking and memory capacity predict social network size," *Social Networks*, 29, 93–104.

Stock, J. H. and M. W. Watson (2005): "Implications of dynamic factor models for VAR analysis," Tech. rep., National Bureau of Economic Research.

Strzalecki, T. (2014): "Depth of Reasoning and Higher-Order Beliefs," *Journal of Economic Behavior and Organization*, 108, 108–122.

Thomson, K. S. and D. M. Oppenheimer (2016): "Investigating an alternate form of the cognitive reflection test," *Judgment and Decision making*, 11, 99.

Vosoughi, S., D. Roy, and S. Aral (2018): "The spread of true and false news online," *Science*, 359, 1146–1151.

Walzer, M. (1973): "Political action: The problem of dirty hands," *Philosophy & public affairs*, 160–180.

Weber, R. A. (2001): "Behavior and Learning in the "Dirty Faces" Game," *Experimental Economics*, 4, 229–242.