Essays on Beliefs and Economic Behaviour



Christopher Roth

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Economics

Hilary 2018

ESSAYS ON BELIEFS AND ECONOMIC BEHAVIOUR

CHRISTOPHER ROTH, KEBLE COLLEGE

DPHIL IN ECONOMICS

HILARY 2018

ABSTRACT

Chapter 1, "Labor Market Concerns and Support for Immigration" (jointly written with Ingar Haaland), examines whether labour market concerns causally affect people's support for immigration. Using a large, representative sample of the US population, we first elicit beliefs about the labour market impact of immigration. To generate exogenous variation in beliefs, we then provide respondents in the treatment group with research evidence showing no adverse labour market impacts of immigration. We find that treated respondents update their beliefs about the labour market impact of immigration and become more supportive of immigration, as measured by self-reported policy views and signatures on real online petitions. We also employ an obfuscated follow-up study which hides the connection between the follow-up and the main study. The treatment effects persist in this setting where experimenter demand is mitigated. Our results demonstrate that beliefs about the labour market impact of immigration are an important causal driver of people's support for immigration.

Chapter 2, "Measuring and Bounding Experimenter Demand Effects" (cowritten with Jonathan de Quidt and Johannes Haushofer), is a methodological paper on experimenter demand effects which proposes the use of "demand treatments" to measure and bound experimenter demand effects. We use a simple Bayesian model to motivate our empirical approach of sending explicit signals of the experimenter's wishes, and conduct a series of large-scale experiments to test our method. We also apply our method to the bounding of treatment effects and structurally estimate a simple model in order to quantify the monetary value of pleasing the experimenter. Our results indicate a limited quantitative importance of demand effects, giving cause for optimism for the validity of experimental results.

In Chapter 3, "Public Debt and the Demand for Government Spending and Taxation" (jointly written with Johannes Wohlfart), we explore how perceptions of government debt shape people's demand for government spending

and taxation, which touches upon the important policy issue of public debt accumulation. To create exogenous variation in people's perceived level of government debt, we provide half of our respondents with information about the debt-to-GDP ratio in the US. We find that most people underestimate the debt-to-GDP ratio and favour a cut in government spending once they learn about the actual amount of debt, but do not alter their attitudes towards taxation. These effects persist in a four-week follow-up and operate through changes in beliefs about fiscal sustainability.

Chapter 4, "Conspicuous Consumption and Peer Effects: Evidence From a Randomised Field Experiment", examines whether peers' consumption choices affect our consumer behaviour. Combining novel Indonesian data on the visibility of consumption goods with a randomised cash transfer program, I find that the expenditure share of visible goods rises for households whose peers' income and visible consumption increase. The results are inconsistent with a signalling model, but consistent with a model of relative concerns over visible goods. Finally, I shed light on the social mechanisms underlying peer effects in consumption.

Supervisor This thesis was supervised by Simon Quinn from the University of Oxford.

Total words This thesis contains approximately 90,000 words.

Acknowledgement of funding I acknowledge funding from the Department of Economics, Keble college Oxford, the Royal Economic Society, the George Webb Medley Fund, the Oxford Economics Papers Fund, the German National Merit Foundation, the German Academic Exchange Service, and the Marie Curie fellowship program. I have also received various research grants from Time-Sharing Experiments for the Social Sciences and the Russel Sage Foundation which were vital for my DPhil research. I am very grateful for the support.

Ethical approval For Chapters 1, 2, 3, and 4 ethical approval has been sought and obtained from the University of Oxford CUREC committee.

Declaration The work in this thesis is based on research I have carried out at the University of Oxford between September 2013 and April 2018. Chapters 1 to 3 of this thesis have not been submitted for a degree at this or any other university. Chapter 4 of this thesis is a substantially revised version of my MPhil thesis.

Christopher Roth (*April 2018)*

Acknowledgements

First, I would like to thank my advisor Simon Quinn. Simon, thanks for your wisdom, your intuition, positivity, brillance and generosity with your time. Thanks for teaching me a great deal on economic research and for always being encouraging about my projects and ideas. Your support is heartfelt and I will forever be grateful to you.

I am also grateful to David Huffman who supervised me during my first PhD year. In particular, I appreciate your insights on what good research is: defining important and interesting questions and then tackling them as convincingly as possible. Moreover, I thank Johannes Abeler for the sharp feedback on some of my attempts to do research. Clement Imbert has also been an incredibly generous source of support when he supervised my MPhil thesis. Thanks for your time and for teaching me a great deal on applied econometrics.

I would like to thank my co-authors who have taught me so much on conducting research: Jon de Quidt, Ernst Fehr, Alexis Grigorieff, Ingar Haaland, Johannes Haushofer, Lukas Hensel, Johannes Hermle, Anselm Rink, Julien Senn, Diego Ubfal and Johannes Wohlfart. I would like to thank Alexis Grigorieff for his input, encouragement, for the great company at the field trips and most importantly his humour. Thanks Alexis for introducing me to experimental economics and behavioural economics. Johannes Haushofer has been a great source of inspiration for me always encouraging me to think big. Moreover, I want to extend special thanks to Johannes Wohlfart for all of the energy he put in our joint projects and for always being open to thinking about new ideas and projects and for igniting my interest in household finance and experimental macro. Jon, thanks for sparking some excitement in economic theory and for inspiring me to be a good writer. Ingar, thanks for a lot of great insights into smart ways to incentivise beliefs about interesting topics, to deal with demand effects, and most importantly for inspiring me to write more efficient code. Lukas, thanks for always being keen to discuss new ideas, for your openness to conduct natural field experiments and for your incredible work ethic, dedication, and stamina in

the face of tight deadlines. Thanks to Johannes Hermle for his brilliant ideas and outstanding technical skills, and his great humour.

My PhD was written at several institutions: Oxford University, Bocconi University (Milan), the Institute on Behavior and Inequality (Bonn), the Busara Center for behavioural economics (Nairobi, Kenya) and the national team for the acceleration of poverty reduction (Jakarta, Indonesia). I thank all members of all of the institutions which allowed me to have a very smooth and fun experience. During my time at Oxford I had a great deal of fun thinking about research and this would not have been possible without the generous help and advice from Stefano Caria, Michael Koelle, Pavel Luengas Sierra, Daniela Scur, Rachel Cassidy, Muhammad Mekki, Jono Lain, Elwyn Davies, Damian Clarke, Markus Parlasca, Marc Witte, Kate Orkin and Simon Unger. I also would like to thank several people I met during my time at the Institute on Behavior and Inequality in Bonn where I learnt an incredible amount on how to design experiments. In particular, I would like to thank Armin Falk, Simon Jaeger, Ulf Zoelitz, Fabian Kosse, Fabian Walders, and Ferderik Schwerter. Special thank goes to Thomas Graeber for many great discussions about research and life and for an enormous amount of useful feedback and laughter.

I have attended various excellent summer schools. I would like to thank Matthew Rabin and David Laibson for their generosity in organising the Russel Sage Summer camp which was a once-in-a-lifetime opportunity to interact with outstanding students and faculty. I would also like to thank Stefano DellaVigna who has been a tremendous source of inspiration and has been extremely supportive ever since I first met him. Moreover, I would like to thank Alexander Cappelen and Bertil Tungodden for generous support on my research projects.

I would like to thank my parents, my grandparents and my brother for their love and unconditional support which made all the difference. Thanks for always being there also in times of trouble. Thanks for listening to me, being encouraging, and for always having an optimistic outlook. Thanks for all the opportunities you provided me with. I would not be where I am today without your love and generosity.

Christopher Roth
Bonn, April 2018

Contents

Introduction **1**

1 Labor Market Concerns and Support for Immigration **4**

1.1 An experiment on beliefs about labour markets 5

1.2 Experimental design and sample 11

1.3 Results 20

1.4 Discussion 40

1.5 Concluding remarks 44

Appendices **46**

1.A Conceptual framework 48

1.B Additional Figures 51

1.C Additional Tables 56

1.D Summary statistics, balance and attrition 68

1.E Beliefs about the survey and the research evidence 78

1.F Miscellaneous 80

2 Measuring and Bounding Experimenter Demand **82**

2.1 Introduction 83

2.2 Theory 91

2.3 Sample and experimental design 101

2.4 Applying the method 107
2.5 Properties of demand effects 122
2.6 Using the method in practice 133
2.7 Conclusion 136

Appendices **138**

2.A Additional Figures and Tables 138
2.B Theoretical Appendix 152
2.C Balance tables, summary statistics, and attrition 189

3 Public Debt and the Demand for Government Spending and Taxation **201**

3.1 An experiment on debt perceptions 202
3.2 Conceptual framework 207
3.3 Experimental design 209
3.4 Beliefs about the debt-to-GDP ratio 216
3.5 The causal effect of information about debt 221
3.6 Mechanisms 235
3.7 Conclusion 242

Appendices **244**

3.A Theoretical appendix 245
3.B Additional Tables 249
3.C Additional Figures 275
3.D Additional results 282

4 Conspicuous Consumption and Peer Effects: Evidence From a Randomised Field Experiment **285**

4.1 Introduction 286

4.2 Conceptual framework 290

4.3 Data 297

4.4 Direct and indirect treatment effects 307

4.5 Mechanisms 311

4.6 Robustness 316

4.7 Conclusion 324

Appendices **325**

4.A Additional Tables 325

4.B Overview of consumption categories 334

Conclusion **337**

References **339**

INTRODUCTION

Standard neoclassical economics posits that economic agents form rational expectations and are perfectly informed. According to this framework, economic agents maximize an objective function based on unbiased beliefs about the world. This framework is very simple and powerful in making sharp predictions about economic behaviour.

The seminal work by Tversky and Kahneman (1974) and Manski (2004) have called into question the assumption that economic agents form rational expectations and are perfectly informed. Tversky and Kahneman (1974) provide striking evidence on people's biases in belief formation and decision-making and Manski (2004) proposes empirical methods to measure people's subjective beliefs to better understand their expectation formation and economic decision-making.

Since people's subjective beliefs are an important determinant of their economic and political behaviour, understanding how people form their beliefs and how they respond to information is relevant for understanding many real world phenomena. For example, people's beliefs are important for determining their voting behaviour. This is particularly relevant for high stakes elections in which politicians deliberately spread "fake news" to win elections. Therefore, it becomes increasingly important to measure people's beliefs and to improve our understanding of how people form their beliefs in the face of new information.

The Brexit vote offers a textbook example for the relevance of biased beliefs about economic facts in shaping voting behaviour. For months, the "Leave campaign" claimed that leaving the EU would allow the UK to save up to 350 millions pounds a week. Their famous slogan was: "We send the EU 350 million a week. Let's fund our NHS instead". Even though this figure was debunked, it was reported to be an important reason for many Britons to vote for "Leave". For example, a post-Brexit poll indicates that two thirds of the "Leave" voters believed that the UK was contributing too much to the EU, and nearly 40% believed that the money given to the EU should be used by the NHS instead.

Can biases in beliefs be corrected, and if so, does information provision affect people's behaviour? Do people's beliefs about the economy affect their political preferences? Do people's beliefs about their peers affect their economic behaviour? This thesis provides evidence on these questions using experimental methods.

Chapter 1 examines the role of labour market concerns in shaping people's immigration preferences. Using a large and representative sample of the US population, we exogenously shift people's beliefs about the labour market of immigration. We do so by providing a random subset of respondents with research evidence showing no adverse labour market impacts of immigration. Thereafter, we measure their immigration preferences using self-reports and a behavioural measure. In contrast to existing correlational evidence, our causal estimates demonstrate that labour market concerns are an important determinant of people's attitudes towards immigration that affects political behavior. There are three aspects of the paper worth noting: First, the paper proposes a novel way of measuring and changing beliefs about people's mental model of the economy by leveraging research evidence from natural experiments. Second, the paper uses real online petitions in order to measure political preferences rather than only hypothetical survey questions. Third, the paper is the first to deal with experimenter demand effects in online experiments through the use of obfuscated follow-up surveys which hide the connection between the follow-up survey and the main study. Thus, the paper makes not only a contribution to the literature on immigration attitudes, but also makes methodological contributions related to the measurement of beliefs and related to the mitigation of experimenter demand effects.

Experimenter demand effects are many times raised as a confound in experiments in general, and more specifically, in experiments that deliberately manipulate beliefs and expectations. Chapter 2 provides systematic evidence on whether people's beliefs about the experimenter's wishes affect people's behaviour in experiments. Motivated by a simple Bayesian model, we employ "demand treatments" to measure and bound experimenter demand effects. Using 19,000 participants we provide evidence on sensitivity to demand effects using 11 classic tasks. We show that people's behaviour responds moderately to their beliefs about the experimenter's expectations. We show how to bound treatment effects and also structurally estimate a simple model. We conclude with recommendations for practitioners.

Chapter 1 examined people's beliefs about the functioning of labour markets and how they affect people's immigration preferences. Similarly, peo-

ple's beliefs about macroeconomic variables may play an important role for their political preferences. Chapter 3 explores how perceptions of government debt shape people's demand for government spending and taxation, which touches upon the important policy issue of public debt accumulation. To create exogenous variation in people's perceived level of government debt, we provide half of our respondents with information about the debt-to-GDP ratio in the US. We find that most people underestimate the debt-to-GDP ratio and favour a cut in government spending once they learn about the actual amount of debt, but do not alter their attitudes towards taxation.

While the previous three chapters have focused on beliefs about facts or beliefs about the experimenter, beliefs and expectations about the behaviour of peers can also be an important determinant of behavior. Chapter 4 studies whether peers' consumption choices affect our consumer behavior. Combining novel Indonesian data on the visibility of consumption goods with a randomised cash transfer program, I find that the expenditure share of visible goods rises for households whose peers' income and visible consumption increase. The results are inconsistent with a signalling model, but consistent with a model of relative concerns over visible goods. I shed light on heterogeneous treatment effects using rich data on social activities at the village level and show that the increases in visible expenditures occur in villages with lower social cohesion.

Taken together, this thesis has a number of unifying themes. First, chapters 1 to 3 provide evidence that there is a large amount of heterogeneity in beliefs. Second, all four chapters show that people's beliefs and expectations can strongly shape their economic decision-making and their behavior. Third, chapters 1 and 2 make methodological contributions related to the mitigation of concerns about experimenter demand effects.

Organisation of the thesis The following four chapters consist of the essays just described, which are presented as self-contained journal-style articles. They all include their own sections on literature review, methodology, data description, results, and discussion. My contributions to existing research are outlined within each chapter.

Chapter 1

Labor Market Concerns and Support for Immigration

Ingar Haaland and Christopher Roth[1]

Abstract We examine whether labour market concerns causally affect people's support for immigration. Using a large, representative sample of the US population, we first elicit beliefs about the labour market impact of immigration. To generate exogenous variation in beliefs, we then provide respondents in the treatment group with research evidence showing no adverse labour market impacts of immigration. We find that treated respondents update their beliefs about the labour market impact of immigration and become more supportive of immigration, as measured by self-reported policy views and signatures on real online petitions. We also employ

[1]Ingar Haaland: Department of Economics, NHH Norwegian School of Economics; Christopher Roth: Department of Economics, University of Oxford. We thank Johannes Abeler, Ned Augenblick, Roland Bénabou, Steve Bond, Alexander Cappelen, David Card, Stefano Caria, Rachel Cassidy, Jack Citrin, Damian Clarke, Vincent Crawford, Jon de Quidt, Dirk Engelmann, Benjamin Enke, Ranveig Falch, Armin Falk, Claudio Ferraz, Thomas Graeber, Alexis Grigorieff, Jens Hainmueller, Johannes Haushofer, Lukas Hensel, Johannes Hermle, Simon Jäger, Michael Kölle, Tom Meling, Edward Miguel, Peter Neary, Muriel Niederle, Markus Parlasca, Ricardo Perez-Truglia, Carlo Pizzinelli, Simon Quinn, Matthew Rabin, Gautam Rao, Anya Samek, Kenneth Scheve, Sebastian Schweighofer-Kodritsch, Frederik Schwerter, Daniela Scur, Stefanie Stantcheva, Eirik Strømland, Adam Szeidl, Erik Sørensen, Bertil Tungodden, Jonas Tungodden, David Yang, Johannes Wohlfart, Florian Zimmermann, several students from Harvard's graduate course in experimental economics, and seminar audiences in Bergen, Berlin, Essex, Lofoten, Munich, New York (NY Fed), Nottingham, Oxford, Richmond (ESA), San Diego, Stanford (SITE), Trondheim, and Warwick for valuable comments. The experiment is registered in the AEA RCT Registry as trial 2247, `https://www.socialscienceregistry.org/trials/2247`. This work was partially supported by the Research Council of Norway through its Centres of Excellence Scheme, FAIR project No 262675. IRB approval was obtained from the University of Oxford. The experimental instructions are available on the following link: `https://goo.gl/SmQmJa`.

an obfuscated follow-up study which hides the connection between the follow-up and the main study. The treatment effects persist in this setting where experimenter demand is mitigated. Our results demonstrate that beliefs about the labour market impact of immigration are an important causal driver of people's support for immigration.

1.1 An experiment on beliefs about labour markets

There is currently a heated political debate about immigration in Western countries. Arguments about the adverse labour market effects of immigrants are prominent in this debate, and voters are deeply divided in their beliefs about the extent to which immigration is good or bad for the economy. In the US, for instance, 67 percent of Republicans say increased immigration has hurt American workers, compared with 30 percent of Democrats.[2] Policy views are equally polarised: 67 percent of Republicans say immigration to the US should be decreased, compared with 33 percent of Democrats.[3] This raises the question whether labour market concerns are a causal driver of people's support for immigration.

A large literature has examined the extent to which labour market concerns affect support for immigration (Card et al., 2012; Hainmueller and Hiscox, 2010; Hainmueller et al., 2015; Sniderman et al., 2004). The current consensus in this literature is that labour market concerns are not a quantitatively important determinant of attitudes towards immigration (Hainmueller and

[2] http://www.pewsocialtrends.org/2016/10/06/1-changes-in-the-american-workplace/, accessed January 25, 2018

[3] http://pewrsr.ch/1Rg1yz7, accessed January 25, 2018

Hopkins, 2014b). A weakness of previous studies, however, is the lack of causal evidence on the link between labour market concerns and support for immigration. We advance this literature by experimentally manipulating beliefs about the labour market impact of immigration. This allows us to provide the first causal evidence on whether people's beliefs about the labour market impact of immigration affect their support for immigration.

In a pre-registered experiment with a large and representative sample of Americans, we first elicit beliefs about the labour market impact of the Mariel boatlift, which is known as the "one historical event that has most shaped how economists view immigration" (Clemens, 2017). During the Mariel boatlift, which was an unexpected mass immigration of Cubans to the United States, the low-skilled workforce in Miami increased by 20 percent over the course of a few months. To generate exogenous variation in beliefs, we provide a random subsample of our respondents with information about research showing no adverse labour market impacts of the Mariel boatlift on wages and unemployment in Miami (Card, 1990). We then measure our respondents' support for immigration using both self-reported attitudes on preferred immigration levels as well as behavioural measures: two anonymous real online petitions proposing changes to the annual cap on visas for low-skilled guest workers to the US. Our use of *anonymous* real online petitions mitigates demand effects by making it impossible for respondents to signal that they conform to the the experimenter's wishes. Finally, we conduct an obfuscated follow-up study one week later in which we hide the connection between the follow-up and the main study. This allows us to further mitigate concerns that treatment effects are biased because of differences in experimenter demand between the treatment and the control group (de Quidt et al., 2018; Zizzo, 2010).

We establish five novel facts about labour market concerns and support for immigration. First, respondents update their beliefs about the labour market impact of immigration when provided with research evidence showing no adverse labour market impacts: the treatment effect corresponds to two-thirds of the difference in beliefs between self-identified Democrats and Republicans. Second, the provision of research evidence increases people's support for low-skilled immigration by 0.14 of a standard deviation. This effect size corresponds to approximately one quarter of the Democrat–Republican difference in policy views. Third, the changes in attitudes also lead to changes in real political behavior: Treated respondents are 2.6 percentage points more likely to sign a real online petition in favour of increasing the annual cap on non-agricultural guest workers to the US. This corresponds to a 66.9 percent increase from the control group mean of 3.9 percentage points ($p<0.01$). We also find a 0.5 percentage point decrease in signatures for a second petition in favour of decreasing the annual cap, compared to a control group mean of 5.5 percentage points, but this difference is not statistically significant ($p=0.27$). Fourth, we demonstrate that labour market concerns are a quantitatively important determinant of attitudes towards immigration: our instrumental variable estimates reveal that a one standard deviation change in perceptions of the labour market impact of immigration changes people's willingness to admit low-skilled immigrants by 0.61 of a standard deviation. Finally, we show that the treatment effects persist over time and remain statistically significant: In the follow-up, treated respondents display 0.09 of a standard deviation more positive attitudes towards low-skilled immigrants, corresponding to 73 percent of the effect size from the main study.

Exploring heterogeneity in treatment responses, we find that respondents

with pessimistic pre-treatment beliefs about the labour market impact of immigration become more optimistic, and vice versa. We confirm these patterns using techniques from machine learning (Athey and Imbens, 2016), which show that people's prior beliefs about the wage impact of the Mariel boatlift are the best predictor of heterogeneous responses to the treatment. Further exploring mechanisms, we find that the treatment does not affect beliefs about the cultural impact of immigration. These results corroborate our interpretation that the treatment operates through changes in beliefs about the economic impact of immigration. Finally, we use data from control group respondents to explore whether cultural or economic factors are more important in explaining attitudes towards immigration. The correlations we uncover suggest that economic concerns are quantitatively more important than cultural concerns.

Our work contributes to several strands of the literature. First, and most closely related, we contribute to the literature on labour market concerns and attitudes towards immigration (Card et al., 2012; Citrin et al., 1997; Dustmann and Preston, 2001, 2006; Facchini et al., 2016, 2009; Hainmueller and Hiscox, 2010; Hainmueller et al., 2015; Hainmueller and Hopkins, 2014a; Iyengar et al., 2013; Mayda, 2006; Scheve and Slaughter, 2001). This literature is mostly based on observational survey data and therefore vulnerable to omitted variable bias and reverse causality (Hainmueller and Hopkins, 2014b).

To overcome the challenges with observational data, researchers have used conjoint experiments which measure support for hypothetical immigrants with randomly assigned characteristics, such as their education levels and whether they plan to find work. These experiments have documented that

Americans' preferences for which immigrants to admit vary little with their own education and labour market position (Hainmueller and Hopkins, 2014a). This has been taken as evidence for a limited role of economic concerns in driving support for immigration, but a weakness of conjoint experiments is that they leave respondents' beliefs about the economic and cultural effects of immigration unrestricted. This makes it hard to shed light on the underlying mechanisms. For instance, the finding in Hainmueller and Hopkins (2014a) that Americans tend to prefer highly educated immigrants that plan to find work could be due to beliefs about their economic or their cultural impact.[4] We employ a different approach by experimentally manipulating our respondents' beliefs about the economic impact of immigration and subsequently measuring immigration preferences with self-reports and behavioural measures. Since our experimental manipulation does not affect our respondents' beliefs about the cultural impact of immigration, we are able to cleanly identify the role of economic concerns in driving support for immigration.

We also contribute to a broader literature that tries to understand the determinants of people's policy preferences by experimentally manipulating beliefs (Alan and Ertac, 2017; Alesina et al., 2018; Cruces et al., 2013; Gilens, 2001; Karadja et al., 2017; Kuklinski et al., 2000; Kuziemko et al., 2015). This literature has mostly focused on correcting biases in people's beliefs about verifiable facts, e.g. related to income inequality. We differ from these studies by shifting people's beliefs about the causal relationship between two economic variables which inform the functioning of markets. Our approach of eliciting people's beliefs about the outcomes of natural experiments allows us to create exogenous variation in beliefs by assigning some respon-

[4] In Section 1.4.3, we provide evidence that people's beliefs about the cultural impact of immigrants depend on the immigrants' economic characteristics, such as their skill levels.

dents to research evidence on the same natural experiments.

In the context of immigration policy, our results differ from previous experiments showing muted responses of policy preferences to information about the fraction of immigrants (Hopkins et al., 2018; Sides and Citrin, 2007), or their characteristics (Grigorieff et al., 2016; Lergetporer et al., 2017).[5] An explanation for why we find stronger responses of policy preferences to new information compared to the previous literature could be that research evidence is more persuasive than are summary statistics.[6]

Our findings also contribute to a long-standing debate in the social sciences which discusses the relative importance of consequential, ideological, and social motives in driving people's political behaviour (Bursztyn et al., 2016; DellaVigna et al., 2017; Downs, 1957; Gerber et al., 2017). Our experimental results support a consequential view on political behaviour by highlighting that changes in beliefs about the economic consequences of a policy can significantly affect political behavior.

Finally, we contribute to the literature on experimental methods to deal with experimenter demand effects (de Quidt et al., 2018; Mummolo and Peterson, 2017; Zizzo, 2010). We employ a follow-up study which is presented as an independent study to participants. Since no treatment is administered in the follow-up study, differential experimenter demand between the treatment and control group can only arise if respondents nonetheless realise that the follow-up is connected to the main study. While follow-up studies are a common tool in information experiments to test whether treatment effects

[5]Our results are also related to online experiments in Japan showing that positive newspaper stories about immigrants can be effective in reducing self-reported anti-immigrant views (Facchini et al., 2016).

[6]In support of this interpretation, we provide new evidence that people's trust in research evidence is higher than their trust in government statistics. Results are available upon request.

persist (Alesina et al., 2018; Grigorieff et al., 2016; Kuziemko et al., 2015), our approach of hiding the connection to the main experiment extends the use of follow-up studies to more directly address concerns about experimenter demand.

The paper proceeds as follows. Section 1.2 describes the experimental design. Section 1.3 presents the main results and explores heterogeneity. Section 1.4 explores mechanisms. Finally, Section 1.5 concludes.

1.2 Experimental design and sample

Our experiment has two parts: A main experiment and an obfuscated follow-up study performed seven days after the main experiment. In the following, we describe the structure of the main experiment and the obfuscated follow-up study (Figure 1.1 provides a summary of the structure).[7]

1.2.1 Main experiment

In the main experiment, we first ask questions about demographics and elicit our respondents' pre-treatment beliefs about the labour market impact of immigration. Then, we expose half of our respondents to the information treatment. Subsequently, we measure our respondents' support for immigration using self-reported policy views and signatures on real online petitions. Finally, we elicit post-treatment beliefs about the labour market

[7]Detailed experimental instructions are available on the following link: `https://goo.gl/g1ABnt`. The Qualtrics survey for the main experiment is available on the following link: `https://nhh.eu.qualtrics.com/jfe/form/SV_8Am0WWUZiq4u2ax`. The Qualtrics survey for the obfuscated follow-up is available on the following link: `https://cessoxford.eu.qualtrics.com/jfe/form/SV_d71YFolo6Dw9Ump`.

impact of immigration.

Pre-treatment beliefs about the Mariel boatlift We first elicit our respondents' beliefs about the labour market impact of the Mariel boatlift, which has strongly shaped how economists view immigration (Clemens, 2017). To familiarise our respondents with the context, we present them with the following text:

> In 1980, Cuba's then President, Fidel Castro, suddenly announced that Cubans wishing to emigrate to the United States were free to do so. This led to an unexpected mass immigration to Miami, Florida, where most of the Cuban immigrants arrived by boat.
>
> With the arrival of the new Cuban immigrants, Miami's workforce grew by 55,000, or 8 percent, almost at once. The new immigrants were mostly low-skilled, which meant that the low-skilled workforce increased by 20 percent.
>
> The large, unexpected addition of 55,000 new immigrants to the Miami workforce has allowed researchers to study the impact of immigration on the labour market. To do so, the researchers studied wage and unemployment changes in Miami after the mass immigration relative to other US cities that, because of geographic distance, were not affected by the mass immigration of Cubans.

Thereafter, we ask our respondents how they think "the mass immigration of Cubans" affected wages and unemployment in Miami for both low-skilled and high-skilled workers. To minimize the risk that respondents

would try to Google the answer, we choose not to incentivize these beliefs[8] and avoided to explicitly mention the Mariel boatlift in the instructions.

Research evidence on the Mariel boatlift Following the belief elicitation, we inform respondents in the treatment group about the results from a seminal study about the labour market consequences of the Mariel boatlift (Card, 1990).[9] Specifically, we present the following text to respondents in the treatment group (see Figure A.11 for a screenshot):

> The researchers who analyzed the short- and long-term effects of the mass immigration of Cubans to Miami concluded that, for both high-skilled and low-skilled workers, the mass immigration had **virtually no effect on wages** and **virtually no effect on unemployment**.
>
> According to the researchers, the mass immigration had virtually no effect on wages and unemployment because the new Cuban immigrants increased the overall demand for goods and services, which created more jobs.

By contrast, respondents in the control group do not receive any information and proceed directly from the belief elicitation questions to the outcome questions.[10]

[8] To test whether incentives would lead to different results, we ran a small study on Amazon Mechanical Turk which showed a qualitatively similar distribution of beliefs (results available upon request).

[9] Card (1990) provides compelling evidence on the labour market consequences of the Mariel boatlift by conducting a difference-in-differences analysis. The findings by Card (1990) that the Mariel boatlift did not have negative labour market impacts are consistent with a large body of economic research on the labour market consequences of immigration (Card, 2005, 2012; Dustmann et al., 2005; Manacorda et al., 2012; Ottaviano and Peri, 2012). However, the ultimate effects of immigration on labour market conditions are still being debated (Borjas, 2014; Card and Peri, 2016).

[10] We did consider including a second treatment with the evidence from Borjas (2017) to

Measuring support for immigration: self-reported policy views To measure how the treatment affects support for immigration, we first investigate self-reported attitudes. We follow Hainmueller et al. (2015) and ask about attitudes towards increasing or decreasing the number of low-skilled (high-skilled) immigrants that are highly familiar (not familiar) with American values and traditions that should be allowed to come and live in the US. All respondents are asked about all four types of immigrants, but we randomize the order of the questions between respondents. For example, to measure our respondents' views on low-skilled immigrants not familiar with American values, we administer the following question:

> Immigrants to the US differ in terms of their professional skill levels as well as their familiarity with American values and traditions. Do you think the US should allow more or less low-skilled immigrants that are not familiar with American values and traditions to come and live here?

For each immigrant type, respondents report their answer on a 5-point Likert scale from (1) "Allow a lot less of these immigrants" to (5) "Allow a lot more of these immigrants" (Section 1.A provides a conceptual framework on how people report their policy views on immigration and how the treatment affects self-reported policy views).

study the effects of negative information, but we ultimately decided to only have one treatment. While Borjas finds that high school drop outs were negatively affected by the Mariel boatlift (Borjas, 2017, but see the critique by Clemens and Hunt (2017)), he supports the general conclusion from Card (1990) that the boatlift had virtually no negative labour market impacts on the majority of the population. The Borjas evidence could therefore make people more optimistic about the labour market impact of immigration if their prior about the overall labour market effect of immigration is pessimistic. We have suggestive evidence that this is the case from a small study on Amazon Mechanical Turk (results available upon request).

Measuring support for immigration: real political behavior After measuring self-reported attitudes, we give our respondents the opportunity to engage in political behaviour using real online petitions as implemented by Grigorieff et al. (2016). We inform our respondents that Congress is debating whether to change the annual cap on non-agricultural guest workers to the US, the H-2B visa program. We chose to focus on the H-2B visa program because it is heavily debated in Congress and because of the tight connection between our informational treatment, which highlighted the labour market impact of low-skilled immigration, and the H-2B visa program, which is a work-related program to bring low-skilled foreign nationals to the US.

In order to make sure that the debate surrounding the H-2B visa program is meaningful to the respondents, we suggest some arguments in favour of both increasing and decreasing the annual cap. Specifically, all respondents receive the following text:

> Some argue that the quota should be increased because private companies say that there are not enough low-skilled American workers for hire. Others argue that the quota should be decreased because access to more foreign workers makes it easier for private companies to cut the wages of low-skilled American workers.

Respondents are then told that they will be given the opportunity to sign one of two petitions related to this debate. The first petition suggests to increase the annual cap from 66,000 to 99,000, whereas the second petition suggests to decrease the annual cap from 66,000 to 33,000. We randomised the order of the petitions between participants. We chose to create petitions with concrete policy proposals to maximize external validity.

We ask our respondents whether they want to sign one of the two petitions. Respondents who say that they want to sign one of the petitions are provided with a link to a real petition that we created on the White House web page, petitions.whitehouse.gov (see Figure A.7 for a screenshot). To identify treatment differences in actual signatures, we provide respondents in the treatment and control group with different links to identical petitions.

Petitions on the White House web page have some noteworthy features. First, petitions are not public until they receive at least 150 signatures, which did not happen for any of the four petitions we created for the purpose of the study. This feature was important to avoid that people outside the experiment signed any of the petitions. Second, the White House requires an email confirmation for petitions to count. This makes signings more costly than on petition pages that do not require any form of verification; respondents have to actually log on to their email account to sign the petition. Third, the petition signatures are anonymous—that is, the experimenter cannot observe the names of the signatures—which strongly mitigates concerns about experimenter demand because respondents cannot use the petitions to signal that they are trying to conform to the experimenter's wishes (Zizzo, 2010).

Post-treatment beliefs about the impact of immigration To explore mechanisms and to confirm that we successfully managed to induce exogenous variation in beliefs, we examine people's perceptions of the economic impact of increased immigration on both "most Americans" and on their own household. In addition to asking respondents about the "overall economic impact," we ask specific questions on wages, unemployment, and taxes. Subsequently, we measure perceptions of the overall impact of immigrants on American culture and society. For this block of questions, we randomize

respondents to answer questions about either the impact of admitting more low-skilled immigrants or admitting more high-skilled immigrants to the US.

1.2.2 Obfuscated follow-up study to deal with demand effects

To mitigate concerns that our treatment effects are biased due to experimenter demand effects, we conduct an obfuscated follow-up survey which hides the connection between the follow-up survey and the main survey.[11] The obfuscation also reduces concerns about differential attrition between the treatment and the control group.

We take several steps to hide the connection between the main study and the follow-up study from the respondents. First, respondents regularly receive invitations to participate in surveys from the market research company we collaborated with. When sending out these invitations, the market research company uses generic invitation forms that only contain information about pay and expected completion time (see Figure A.8 for a screenshot). Second, we use different consent forms for the two studies: In the first study, respondents are forwarded to a survey with a consent form from the Norwegian School of Economics; in the second, with a consent form from the University of Oxford (see Figure A.9 and Figure A.10 for screenshots). We also use different layouts for the two surveys. Third, to make the follow-up seem like an independent study, we first ask respondents a series of questions about their demographics. Fourth, to further obfuscate the purpose of

[11] The actual number of days between the main study and the follow-up study varied between one and fourteen days for all subjects, but the average difference was seven days.

the follow-up study, we ask several questions about government spending, taxation, and redistribution before we ask any questions about immigration.

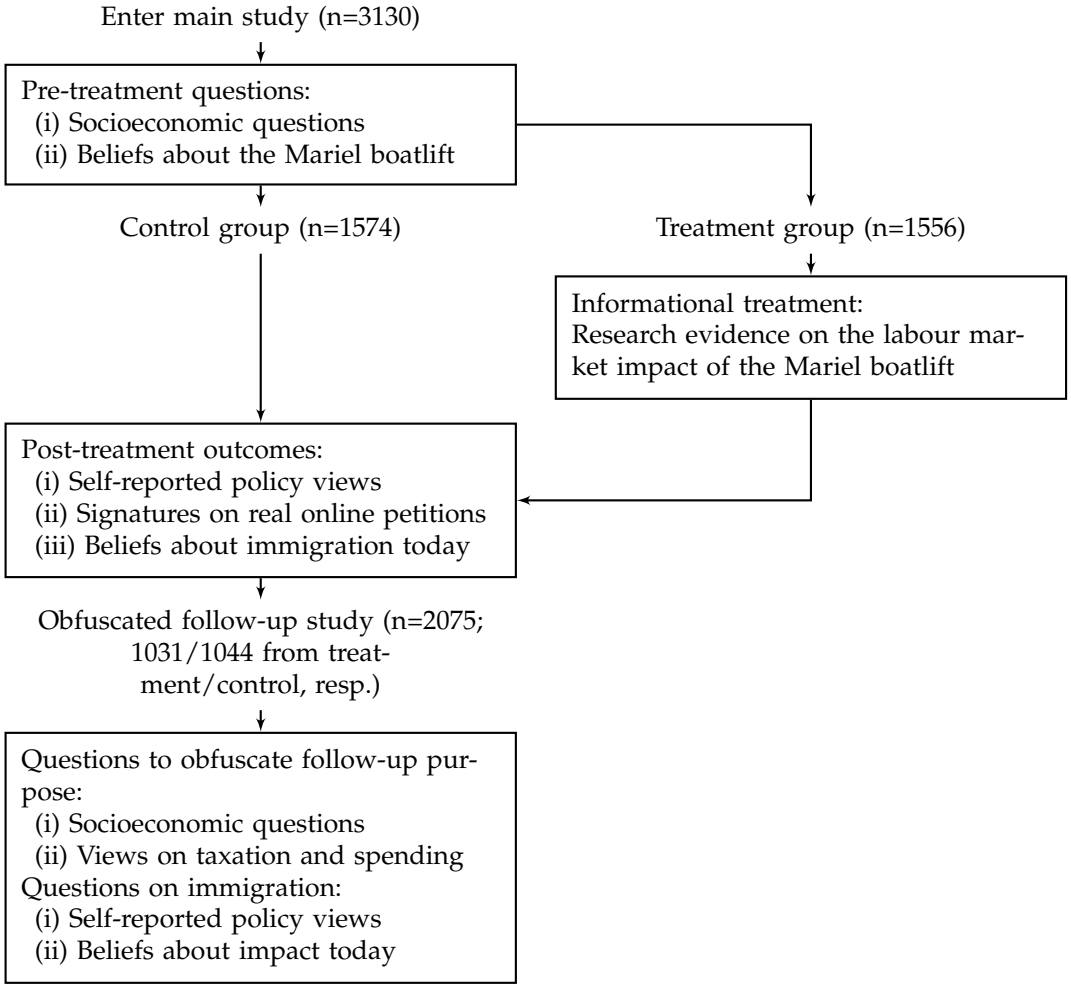
At the end of the follow-up study, we ask three key outcome questions about immigration. Since three questions about immigration may send a signal that we are interested in immigration (and thus increase the chance that respondents realize that the two studies are connected), we ask each question on a separate page with the most important outcome question (preference for low-skilled immigration) on the first of these three pages. To further minimize the chance that respondents realize the relationship between the two studies, we use different wordings for the questions on immigration in the follow-up compared to the main study.[12] This mitigates the concern that treatment effects persist because of people's taste for consistency in survey responses (Falk and Zimmermann, 2012).

1.2.3 Sample

We recruited respondents using Research Now, an online market research company in the US that is regularly used by researchers to conduct academic studies. We recruited 3130 respondents who are broadly representative of the US population in terms of gender, age, income, race, and regions. All respondents who had finished the main study were invited to participate in the obfuscated follow-up study, for which we received 2075 respondents. The experiment was run in late May and early June 2017. We sub-

[12]For instance, in the main study, we ask respondents the following question: "Do you think the US should allow more or less low-skilled immigrants that are not familiar with American values and traditions to come and live here?" In the follow-up study, we drop the distinction about familiarity with American values and traditions and simply ask: "In your view, should immigration of workers with little to no education be kept at its present level, increased, or decreased?"

Figure 1.1: Overview of the experiment



mitted a pre-analysis plan to the AEA RCT Registry before we collected any data for the experiment. We pre-specified the sample size, empirical specifications, and our hypotheses in a pre-analysis plan available on the following link: `https://www.socialscienceregistry.org/trials/2247`.

1.2.4 Summary statistics and balance

In the main experiment, the median age is 45; the median income is $62,500; 50 percent of respondents have at least a two-year college degree; 48 percent are male; and 45 percent are in full-time work (see Table A.15). Both our

main sample and the follow-up sample are representative of the US population in terms of age, region, gender, and household income (see Table A.4). Treatment and control group are balanced in terms of observables both in the main study and the follow-up (see Tables A.16 and A.17) and there is no differential attrition in the response rates to the follow-up (see Table A.18). Since there is some selective attrition unrelated to the treatment status, but related to some demographics (for details see Table A.18), we include controls in our specifications analyzing the follow-up results.

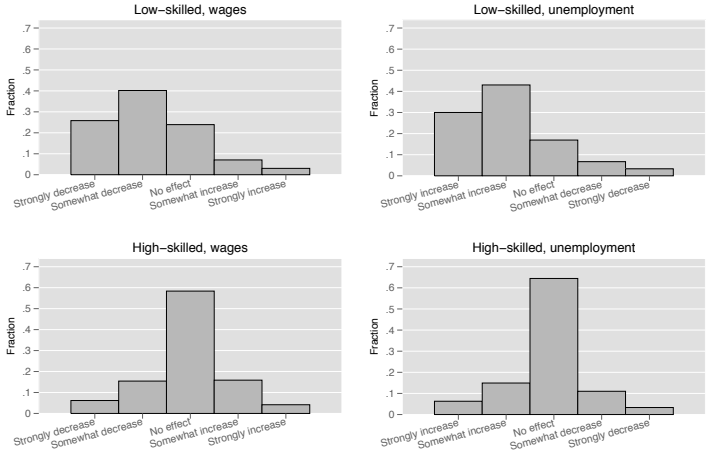
1.3 Results

The main results presented in this section were pre-specified in the analysis plan. We also report some results that were not pre-specified, but which naturally follow from the pre-analysis plan. This includes the results presented in Section 1.3.2 on belief updating and the IV results in Section 1.3.5.

1.3.1 Pre-treatment beliefs about the Mariel boatlift

To aid interpretation of our main results, it is informative to first investigate which pre-treatment beliefs our respondents hold about the labour market impact of the Mariel boatlift. Figure 1.2 shows the distribution of these beliefs. For low-skilled workers in Miami, the large majority of our respondents thinks that the boatlift negatively affected wages and unemployment. By contrast, for high-skilled workers in Miami, the large majority of our respondents thinks the boatlift had no effect on wages and unemployment. This suggests that the treatment has more scope to change beliefs about the impact of low-skilled immigration than high-skilled immigration.

Figure 1.2: Pre-treatment beliefs about the labour Market impacts of the Mariel boatlift



Notes: The figure shows the distribution of pre-treatment beliefs about the labour market impacts of the Mariel boatlift on wages (left panels) and unemployment (right panels) for low-skilled workers (top panels) and high-skilled workers (bottom panels). Respondents are asked two question on wages: "In the five-year period after 1980, how do you think wages of low-skilled (high-skilled) workers in Miami were affected by the mass immigration of Cubans?" The responses are on a 5 point-scale ranging from (1) strongly decrease to (5) strongly increase. On top of that respondents answer two questions on unemployment: "In the five-year period after 1980, how do you think unemployment among low-skilled (high-skilled) workers in Miami was affected by the mass immigration of Cubans?" The responses are on a 5 point-scale ranging from (1) strongly increase to (5) strongly decrease.

Do people with different demographic characteristics hold different beliefs about the labour market impact of immigration? We find that our respondents' pre-treatment beliefs about the Mariel boatlift vary systematically by their background characteristics (see Figure A.2). Differences in beliefs between self-identified Republicans and Democrats are especially pronounced. For beliefs about the impact on low-skilled wages in Miami, the gap between Republicans and Democrats corresponds to approximately one quarter of a standard deviation. We also observe significant correlations between pre-treatment beliefs and college education, race, work status, age, and income. We find it reassuring that our pre-treatment beliefs are meaningfully correlated with key demographics and political views as it suggests that our respondents paid careful attention to the questions.

1.3.2 Do beliefs respond to the treatment?

To investigate whether the treatment successfully affects our respondents' beliefs about the labour market impact of immigration, we estimate the following equation using OLS:[13]

$$y_i = \alpha_0 + \alpha_1 \mathrm{T}_i + \boldsymbol{\alpha}_2 \mathbf{x}_i + \varepsilon_i$$

where y_i is the outcome of interest; T_i is an indicator for whether subject i received the research evidence; $\mathbf{x}_i$ is a vector of pre-specified controls[14]; and ε_i is an individual-specific error term. For all specifications, we use robust standard errors.

Table 1.1 presents regression results for our main outcome measures on post-treatment beliefs about the labour market impact of immigration. We elicit these beliefs by asking our respondents how they think increased immigration to the US "would affect most Americans over the next five years". In the regressions, we z-score all outcomes using the mean and standard deviation of the control group. Columns 1 to 5 show our first main result:

Result 1. *People develop more optimistic beliefs about the labour market impact of low-skilled and high-skilled immigration in response to the research evidence. The treatment closes up to two-thirds of the Democrat-Republican difference in beliefs.*

For instance, the treatment increases optimism about the wage impact of

[13] All of the results in this paper are robust to employing ordered response models instead of OLS (results available upon requests).

[14] The controls, which were pre-specified, include gender, age, ethnicity, region, household size, household income, education, employment status, party affiliation, whether the respondent was born in the US, whether the subject's parents were born in the US, self-perceived skill-level, and pre-treatment beliefs about the labour market impact of low-skilled (or high-skilled) immigration. The controls are coded as described in the pre-analysis plan. As pre-specified, we also report results without controls in the tables.

low-skilled immigration by 0.16 of a standard deviation ($p < 0.01$). This effect size corresponds to two-thirds of the difference in beliefs between self-identified Democrats and Republicans. We observe effects sizes of similar magnitude for beliefs about high-skilled immigrants (columns 3 and 4). This variation in beliefs about the labour market impact of immigration allows us to investigate whether changes in beliefs have a causal impact on support for immigration.

Table 1.1: Beliefs about the labour market impact of immigration (post-treatment)

	Low Skilled		High Skilled		Follow-up
	Wages	Employment	Wages	Employment	Overall Labor Concerns
Treatment	0.178***	0.148***	0.170***	0.241***	0.115***
	(0.046)	(0.047)	(0.047)	(0.047)	(0.040)
Observations	1489	1484	1487	1481	2075

Notes: The table shows OLS regression results where the dependent variables are beliefs about the labour market impact of immigration. In the main study, respondents were asked how they thought admitting more low-skilled/high-skilled immigrants would affect wages and unemployment for "most Americans". We randomised whether respondents answered these questions for low-skilled or high-skilled immigrants. In the follow-up, respondents were asked whether they thought increased immigration would hurt American workers. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views on the effect of immigration. The outcome variables are z-scored using the mean and standard deviation in the control group. Controls include gender, age, ethnicity, region, household size, household income, education, employment status, party affiliation, whether the respondent was born in the US, whether the subject's parents were born in the US, self-perceived skill-level, and pre-treatment beliefs about the labour market impact of low-skilled (or high-skilled) immigration and are coded as described in the pre-analysis plan. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses.

1.3.3 Do policy preferences respond to changes in beliefs?

Table 1.2 presents regression results for our main outcome measures on self-reported policy views. We use a pre-specified index to assess self-reported policy views on low-skilled immigration. The index is defined as the average of the standardised responses to the following two questions: support

for increasing the number of low-skilled immigrants that are (*i*) "highly familiar" with American values and traditions and (*ii*) "not familiar" with American values and traditions. Column 3 shows our second main result:

Result 2. *The treatment increases support for low-skilled immigration by 0.14 of a standard deviation (p<0.01). This effect size corresponds to one fifth of the Democrat–Republican gap in policy preferences.*

Do treatment effects depend on the familiarity of immigrants with American values and traditions? Column 1 shows that the treatment increases support for low-skilled immigrants that are not familiar with American values and traditions, which is the type of immigrants that most strongly resembles the Cubans from the Mariel boatlift, by 0.17 of a standard deviation.[15] This effect size is larger than the 0.12 standard deviation change in support for low-skilled immigrants that are highly familiar with American values and traditions (p=0.056).

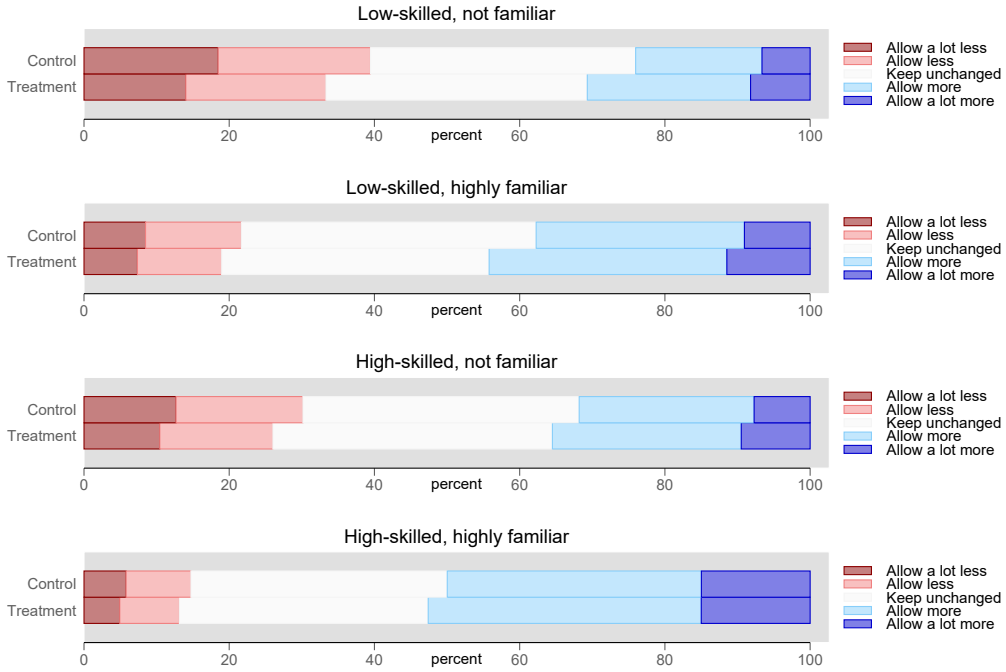
Does the treatment mainly increase support for immigration or mainly decrease opposition to immigration? Figure 1.3 which shows the distribution of responses for both control and treatment group respondents, highlights that the treatment both makes people less likely to support decreased immigration and more likely to support increased immigration. For instance, for low-skilled immigrants that are not familiar with American values and traditions, the treatment increases the share saying they want to "allow more" or "allow a lot more" of these immigrants by 21.7 percent and decreases the share saying they want to "allow less" or "allow a lot less" by 15.6 percent.

We also observe a significant treatment effect on support for high-skilled

[15] At the end of the survey, 77 percent of our respondents said that they thought that the Cuban immigrants from the Mariel boatlift were not familiar with American values and traditions.

immigrants (columns 4–6 of Table 1.2), but the effect sizes are smaller than for low-skilled immigrants ($p<0.01$). It is not surprising that we observe larger treatment effects on low-skilled immigrants as the research evidence is about Cuban immigrants, who are described as mostly low-skilled in our experimental instructions.

Figure 1.3: Distribution of attitudes towards immigrants: Treatment vs. control



Notes: The figure shows the distribution of attitudes towards low-skilled/high-skilled immigrants that are highly familiar/not familiar with American values and traditions, disaggregated by the treatment and control group.

Table 1.2: Self-reported attitudes towards immigration (post-treatment)

	Low-skilled (Main Study)			High-skilled (Main Study)			Follow-up	
	Not familiar	Familiar	Index	Not familiar	Familiar	Index	Low-skilled	High-skilled
Panel A: With controls								
Treatment	0.169***	0.117***	0.143***	0.105***	0.045	0.075**	0.095**	0.059
	(0.033)	(0.034)	(0.030)	(0.033)	(0.034)	(0.030)	(0.040)	(0.042)
Adjusted p-value	[0.001]	[0.001]		[0.003]	[0.104]			
Observations	3130	3130	3130	3130	3130	3130	2075	2075
Panel B: Without controls								
Treatment	0.165***	0.122***	0.144***	0.106***	0.049	0.078**	0.103**	0.062
	(0.036)	(0.036)	(0.033)	(0.036)	(0.035)	(0.032)	(0.043)	(0.043)
Adjusted p-value	[0.001]	[0.001]		[0.005]	[0.089]			
Observations	3130	3130	3130	3130	3130	3130	2075	2075

Notes: The table shows OLS regression results where the dependent variables are attitudes to the different types of immigrants. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. The indices are created by taking the mean of the responses to immigrants with different familiarity with American values and traditions for each skill level. Adjusted p-values are in brackets. "Treatment" is an indicator equal to 1 if respondents received the research evidence. Controls are coded as described in the pre-analysis plan and are listed in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

1.3.4 Does people's political behaviour respond to changes in beliefs?

To examine whether changes in self-reported attitudes also translate into changes in actual political behavior, we analyze whether the treatment affects people's willingness to sign real online petitions regarding the cap on non-agricultural guest workers to the US, the H-2B visa program.

We first analyze how the treatment affects respondents' stated willingness to sign one of the petitions. Table 1.3 presents the results.[16] Column 1 shows that the treatment increases the share of respondents who say they would sign the petition to increase the annual cap on H-2B visas by 4.4 percentage points ($p<0.01$), which corresponds to a 15.3 percent increase from the control group mean of 28.6 percentage points. Similarly, column 2 shows that the treatment decreases the share who say they want to sign the petition to decrease the annual gap by 6.1 percentage points ($p<0.01$), corresponding to a 19 percent decrease from the control group mean of 31.9 percentage points. An important question is whether the observed changes in intentions to sign the petitions are reflected in actual petition signatures. Column 4 demonstrates that this is the case for the petition suggesting an increase in the annual cap:

Result 3. *The treatment increases the share actually signing the petition in favour of increasing the annual cap by 2.6 percentage points ($p<0.01$). This corresponds to a 66.9 percent increase from the control group mean of 3.9 percentage points.*

Column 5 shows that the treatment decreases the share actually signing

[16] In this section, we focus on the results without controls since controls are not available for actual signings of the petitions.

the petition in favour of reducing the annual cap by 0.5 percentage points, which corresponds to a 9 percent decrease in signatures from the control group mean of 5.5 percentage points. This difference, however, is not statistically significant ($p>0.10$). One possibility as to why intentions and actual behaviour differ for this petition could be that the treatment was less effective in changing beliefs for respondents at the margin of actually signing the petition than for respondents at the margin of intending to sign the petition.[17]

Our results illustrate the importance of collecting data on real behaviour in addition to intentions: We find large and significant treatment effects on intentions to sign both petitions, but comparing actual signatures we only find large and significant treatment effects on one of the petitions. Overall, we find it encouraging that the treatment effect carried over from intentions to actual behaviour on the petition suggesting to increase the annual cap on H-2B visas. This result demonstrates a causal role of labour market concerns on people's political behavior. Furthermore, we believe concerns about demand effects are strongly mitigated for the petitions not only because they are measures of real behaviour but also because the signings are anonymous. Anonymity is important because it implies that respondents cannot signal that they conform to the experimenter's wishes.

[17]For a discussion on how treatments effects in information experiments depend on the density of respondents at the margin of taking an action and how much the beliefs of those at the margin are moved, see Coffman et al. (2015).

Table 1.3: Online petitions on H-2B visas (post-treatment)

	Intention: H2B Visas			Actual signatures H2B Visas	
	Increase	Decrease	Net support	Increase	Decrease
Panel A: With controls					
Treatment	0.046***	-0.058***	0.134***	–	–
	(0.016)	(0.016)	(0.034)	–	–
Observations	3130	3130	3130	–	–
Panel B: Without controls					
Treatment	0.045***	-0.060***	0.139***	0.026***	-0.005
	(0.016)	(0.016)	(0.036)	(0.007)	(0.007)
Control Mean	0.285	0.318	-0.000	0.039	0.055
Observations	3130	3130	3130	3130	3130

Notes: The three first columns show regression results where the dependent variable is intention to sign the petitions. "Increase" ("Decrease") is an indicator equal to 1 if a respondent wanted to sign the petition suggesting to increase (decrease) the annual cap on the H-2B visa program. "Net support" is a z-scored transformation of a variable taking value 1 (-1) if a respondent wanted to sign the petition to increase (decrease) the annual cap on the H-2B visa program and 0 otherwise. The two last columns show actual signatures. Since we only observe actual signatures on the treatment group level, we cannot include controls and run regressions for these outcomes. To do testing, we calculate standard errors using the standard formula for proportion tests. "Treatment" is an indicator equal to 1 if respondents received information about the labour market impact of the Mariel boatlift. We use the same controls as in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. For the questions on intention to sign the petitions, we apply robust standard errors in parentheses.

1.3.5 Quantifying the importance of labour market concerns

After establishing that our research evidence changes self-reported attitudes and political behavior, we shed further light on the quantitative importance of labour market concerns. To do so, we examine how a given change in beliefs about the labour market impact of immigration affects people's policy views by using our experimentally induced variation in beliefs as an instrument.

We provide evidence that the the experimental variation is both a valid in-

strument and that it did not affect beliefs unrelated to labour market concerns. Specifically, the informational treatment shifted our respondents' beliefs about the impact on wages and employment, labour_concerns$_i$[18], but did not affect beliefs about the impact of immigration on culture, or public finances (see Table A.1).[19] labour_concerns$_i$ is defined as the unweighted index of people's beliefs about the labour market impact of immigration on *(i)* the wages of most Americans, and *(ii)* the employment among most Americans. In the IV regression, we instrument this index with the treatment indicator, T_i as follows:

$$y_i = \beta_0 + \beta_1 \widehat{\text{labour_concerns}}_i + \boldsymbol{\beta}_2 \mathbf{x}_i + \varepsilon_i$$

where

$$\text{labour_concerns}_i = \gamma_0 + \gamma_1 \mathrm{T}_i + \boldsymbol{\gamma}_2 \mathbf{x}_i + \epsilon_i$$

The controls, $\mathbf{x}_i$, are included and coded as described in the pre-analysis plan. Panel A of Table 1.4 reports the results of an IV regression where beliefs about the labour market impact of low-skilled immigration on "most Americans" are instrumented by the treatment indicator.

Result 4. *We find that a one standard deviation change in beliefs about the economic impact of low-skilled immigrants translates into a 0.61 standard deviation change in attitudes towards these immigrants ($p<0.01$).*

These results demonstrate that group-level concerns about how low-skilled

[18]The implied first-stage F-Stat is for beliefs about the impact of low-skilled immigrants and high-skilled immigrants is above 20, which suggests that our instrument is sufficiently informative.

[19]The underlying identifying assumption is that the new information has no effect on attitudes towards immigrants except through beliefs about their impact on the labour market. This rules out that people develop more positive feelings towards immigrants as a result of the research evidence. The persistence of treatment effects presented in section 1.3.6, as documented in the follow-up study, indicates that the treatment predominantly operates through changes in beliefs.

immigration affects most Americans are a quantitatively important driver of attitudes towards immigration. We also present the IV results for attitudes towards high-skilled workers. A one standard deviation change in beliefs about the labour market impact of high-skilled immigrants translates into a 0.40 standard deviation change in attitudes towards these immigrants ($p<0.01$).

As a robustness check we also demonstrate that the IV results are barely affected by controlling for people's beliefs about the impact of immigration on wages and employment of the own household, as well as beliefs about the tax burden created by immigrants for most Americans and the own household, as well as beliefs about the cultural impact of immigration (see Panel B of Table 1.4).

How externally valid are our IV estimates? To answer this question, we compare our IV estimates to the corresponding OLS estimates which are shown in Panels C and D of Table 1.4. The OLS estimates are smaller than the IV estimates (about one third of the effect size), which suggests that the confounding factors in the OLS regressions downward bias the estimated relationship between labour market concerns and support for immigration. One possible explanation for this downward bias could be measurement error in beliefs. Moreover, the relatively large effect size could also be explained by the fact that the IV estimates capture the local average treatment effect for the compliers, i.e. the subsample of respondents who change their beliefs about the overall economic impact of immigration. It is possible that respondents who change their beliefs in response to new information may have a weaker ideological stance on immigration, and may also be more willing to change their policy preferences.

Table 1.4: The causal effect of labour market concerns: Subsample of respondents with pessimistic priors

	Low-skilled			High-skilled		
	Not familiar	Familiar	Index	Not familiar	Familiar	Index
Panel A: IV						
Labor Market Concerns	0.722***	0.512**	0.617***	0.546***	0.250	0.398**
	(0.227)	(0.238)	(0.210)	(0.180)	(0.187)	(0.163)
Adjusted p-value	[0.004]	[0.010]	[0.004]	[0.007]	[0.064]	[0.014]
Observations	1221	1221	1221	1225	1227	1227
Panel B: IV with additional controls						
Labor Market Concerns	0.789***	0.585**	0.687***	0.623*	-0.025	0.299
	(0.268)	(0.287)	(0.250)	(0.349)	(0.360)	(0.310)
Adjusted p-value	[0.009]	[0.013]	[0.009]	[0.286]	[1.000]	[0.503]
Observations	1221	1221	1221	1225	1227	1227
Panel C: OLS						
Labor Market Concerns	0.387***	0.292***	0.292***	0.362***	0.377***	0.377***
	(0.032)	(0.035)	(0.035)	(0.032)	(0.033)	(0.033)
Adjusted p-value	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Observations	1221	1221	1221	1225	1227	1227
Panel D: OLS with additional controls						
Labor Market Concerns	0.290***	0.223***	0.292***	0.227***	0.200***	0.200***
	(0.040)	(0.044)	(0.035)	(0.045)	(0.048)	(0.048)
Adjusted p-value	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Observations	1221	1221	1221	1225	1227	1227

Notes: These regressions are based on the subsample of respondents who thought that the Mariel boatlift had either negative or no effects on wages and unemployment. The dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants." Moreover, we construct an index for attitudes towards low-skilled immigrants and towards high-skilled immigrants. "Labor Market Concerns" is an unweighted index of people's beliefs about the effect of low-skilled or high-skilled immigration on (i) "wages for most Americans" and (ii) "job opportunities or job security for most Americans." In Panels A and B, we instrument "Labor Market Concerns" with the treatment indicator. In Panels C and D, we run OLS regressions. We randomised whether people were asked about the impact of low-skilled or high-skilled immigrants. All variables have been normalised. All regressions include the controls described in Table 1.1. Panels B and D additionally control for people's beliefs about the impact of immigration on wages and employment of the own household, as well as beliefs about the tax burden created by immigrants for most Americans and the own household, as well as beliefs about the cultural impact of immigration. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

1.3.6 Do changes in attitudes persist in the obfuscated follow-up?

There are three main concerns about the evidence from the main experiment. First, treatment effects may be short-lived and may not reflect genuine changes in beliefs, but rather emotional responses. Second, the treatment effects on self-reported policy views may be biased due to differences in experimenter demand effects across treatment arms. That is, respondents in the treatment group may infer from the research evidence that the experimenter has pro-immigrant attitudes and, and as a result, report more positive views on immigration in order to signal that they are trying to please the experimenter.[20] Third, it is possible that salience of the economic effects of immigration induced by the elicitation of beliefs on the labour market effects of the Mariel boatlift led respondents to neglect other aspects of immigration, such as its impact on culture.

The obfuscated follow-up mitigates these concerns. First, persistence of treatment effects over time would indicate that respondents genuinely updated their beliefs about the labour market impact of immigration. Second, since no treatment was administered in the follow-up study, differential experimenter demand between the treatment and control group can only arise if respondents nonetheless realize that the follow-up is connected to the main study. Third, since the economic impact of immigration was not made salient in the follow-up, treatment effects are unlikely to be inflated because respondents only focused on the economic impact of immigration.

[20]This concern is mitigated for the behavioural measures as the signings of the petitions were anonymous, making it difficult for the respondents to use the petitions to signal that they are trying to please the experimenter. Moreover, recent evidence suggests that demand effects are quantitatively not important (de Quidt et al., 2018; Mummolo and Peterson, 2017).

Columns 7 and 8 of Table 1.2 show our fifth main result:

Result 5. *The treatment effect on support for low-skilled immigration persists in the obfuscated follow-up. The point estimate of 0.09 of a standard deviation ($p<0.05$) corresponds to 73 percent of the effect size observed for the same respondents in the main experiment.*[21]

That we observe a somewhat lower effect size on support for low-skilled immigration in the follow-up could reflect some forgetting of the information, but the effect size in the follow-up is not significantly different from the effect size observed in the main study (p=0.30). For high-skilled immigrants, we observe an estimate of 0.057 of a standard deviation in the follow-up (p=0.17), corresponding to 70 percent of the effect size from the main experiment. This is not significantly different from the effect size observed in the main study (p=0.74). Column 5 of Table 1.1 also shows persistence in changes in beliefs about the overall labour market impact of immigration ($p<0.01$).[22] This suggests that our results reflect genuine belief updating and that demand effects and salience were not the main drivers of the treatment effects on self-reported policy views observed in the main experiment.

To further strengthen our claim that demand effects were not driving our main results, we conducted an experiment to test whether follow-up studies are an effective tool to mitigate demand effects. In the experiment, which was conducted on Amazon Mechanical Turk, we examine the responsiveness of immigration policy views in response to a 'demand treatment' as proposed by de Quidt et al. (2018). Specifically, we tell some respondents

[21] Results on treatment effects for the sample of respondents who also complete the follow-experiment can be seen in Table A.3.

[22] We explored whether there was any heterogeneous treatment responses by number of days between the main study and the follow-up, but did not find any systematic differences based on this measure.

that "we are a group of researchers in favour of immigration," while the remaining respondents receive no such message. While this message increases people's support for immigration instantaneously after receiving the 'demand treatment,' our treated respondents' self-reported support for immigration is essentially unchanged in a one-week follow-up (see Table A.13). The treatment effect observed in the one-week follow-up is also significantly different from the treatment effect immediately after receiving the demand treatment ($p<0.05$). Since we consider the deliberately induced experimenter demand as an upper bound for the demand that our research evidence may induce, this suggests that follow-up surveys are an effective tool for dealing with demand effects.[23]

1.3.7 Heterogeneous effects

Does people's belief updating depend on their prior beliefs? We assess the heterogeneous effects by prior beliefs, using the following specification:

$$y_i = \delta_0 + \delta_1 \mathrm{T}_i + \delta_2 \mathrm{prior}_i + \delta_3 \mathrm{T}_i \times \mathrm{prior}_i + \boldsymbol{\delta}_4 \mathbf{x}_i + \varepsilon_i$$

where the interaction term, prior_i, is pre-treatment standardised beliefs about the labour market impact of immigration.[24] People's belief updating in response to the research evidence should strongly depend on their priors about how immigrants affect local labour markets. In what follows, we explore whether beliefs about the labour market impact of low-skilled immigrants on most Americans depend on our respondents' pre-treatment be-

[23] This complements evidence by Cavallo et al. (2016) who show that follow-ups can mitigate concerns about numerical anchoring and spurious learning.

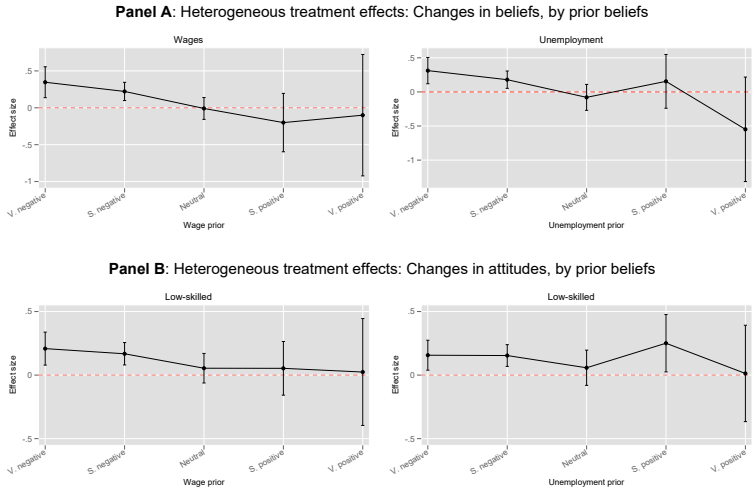
[24] We construct this belief measure by creating an unweighted index of people's standardised beliefs about the effect of immigration on (i) wages and (ii) unemployment.

liefs about the effects of the Mariel boatlift.

Individuals with pessimistic beliefs to begin with strongly update their beliefs about the labour market impact of immigrants on most Americans (see Figures 1.4 and A.4). On the other hand, respondents with priors that are aligned with the research evidence do not update their beliefs. Respondents who thought that the Mariel boatlift had positive labour market effects, develop more pessimistic views about the labour market impact of immigration. However, these patterns are noisily measured and not statistically significantly different from zero (see Table A.19).[25] We previously uncovered a large degree of heterogeneity in preferences between Republicans and Democrats and high-skilled vs. low-skilled respondents. Do these demographics also predict people's response to the treatment information? Republicans do not respond differently to the treatment compared to non-Republicans (see Panel B of Tables A.19 and A.20). This could reflect two opposing factors: On the one hand, Republicans are more likely to dismiss the evidence from the research study (see Table A.26); on the other hand, there is more scope for increases in support for immigration for Republicans as they have more negative views to begin with. Finally, factor proportion models predict that low-skilled workers lose from low-skilled immigration and gain from high-skilled immigration. If people's self-interested labour market concerns mattered for their attitudes towards immigrants, we would expect to see larger treatment effects on attitudes towards low-skilled (high-skilled) immigrants for low-skilled (high-skilled) workers compared to high-skilled (low-skilled) workers. We find no significant treatment heterogeneity by skill level (see Panel C of Tables A.19 and A.20).

[25] This could be due to the fact that pre-treatment beliefs are correlated with other covariates (such as political affiliation, attention or education) for which we may expect patterns of heterogeneity going in the opposite direction.

Figure 1.4: Heterogeneous treatment effects



Notes: The figure shows treatment effects depending on people's prior beliefs about the wage impact and the employment effect of the Mariel boatlift on low-skilled Americans. In Panel A we show results on an index of people's beliefs about the labour market impact of immigration today. In Panel B we display results on an index of attitudes towards increase/decreasing low-skilled immigration.

Machine-Learning Approaches to Heterogeneity The selection of pre-specified covariates for the heterogeneity analysis in the previous section was motivated by theoretical considerations. However, this analysis does not allow us to assess which of all of the covariates contained by our dataset are the most relevant ones for shaping heterogeneous responses to our information treatment. To answer this question, we employ newly developed tools from machine learning (and more specifically Classification and Regression Trees (CART)). This allows us to select those variables which predict maximum differences in the magnitude of treatment effects for different subpopulations (Athey and Imbens, 2016), while simultaneously controlling for multiple hypothesis testing. The algorithm constructs a tree by sequentially dividing the dataset into two groups with the target of minimizing the mean squared error of the estimated treatment effects. To deal with over-fitting,

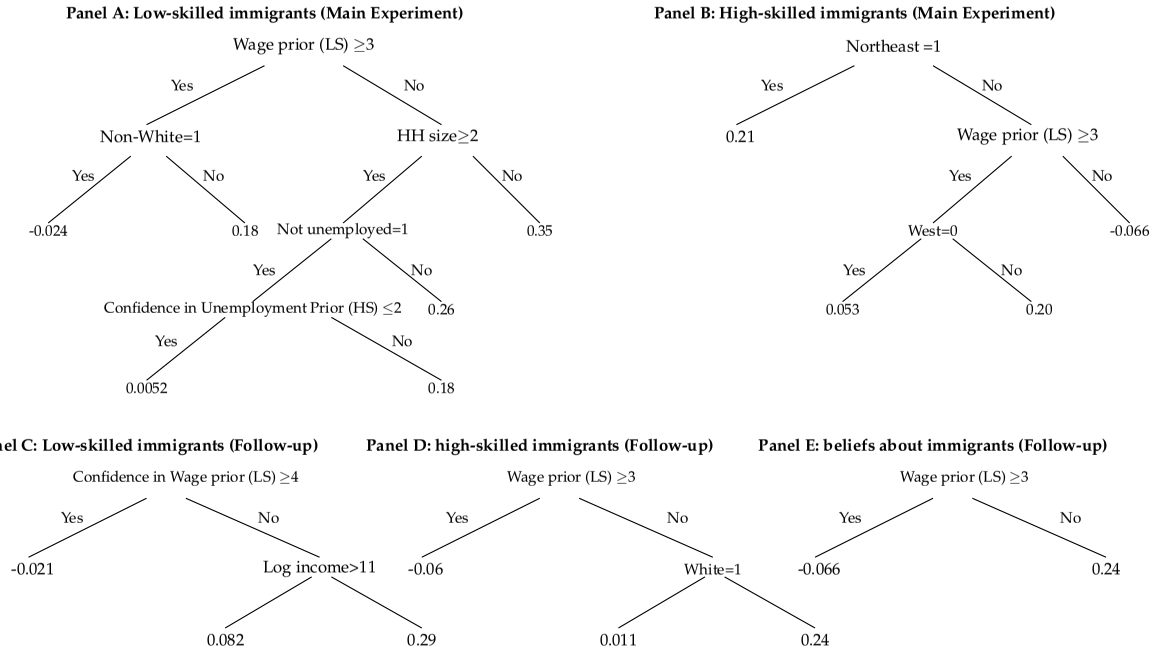
the algorithm uses a cross-validation set to reduce the depth of the tree.[26] Figure 1.5 presents the causal trees obtained by the CART algorithm for the outcome variables of interest. Panels A and B show results for an index of people's attitudes towards low-skilled immigrants and high-skilled immigrants from the main experiment, respectively. Panels C to E display the causal trees for people's immigration preferences and beliefs as measured in the obfuscated follow-up survey. The causal trees reveal our sixth main result:

Result 6. *Respondents' prior beliefs about the wage impacts of the Mariel boatlift are most powerful at predicting differences in responses to the information treatment both in terms of immigration preferences and beliefs about the labour market impact of immigration.*

The trees show that people's wage prior or their confidence in the wage prior appear at the top of almost all of the causal trees presented in Figure 1.5. Specifically, the trees confirm that respondents with optimistic prior beliefs about the Mariel boatlift mostly develop more negative attitudes towards immigrants, consistent with genuine belief updating. At the lower end of the causal trees both proxies for economic status (such as income and employment status), but also indicators of racial group membership appear in the tree. This suggests that both economic and cultural considerations shape people's responsiveness to the information treatment. Taken together it is reassuring, that the data-driven machine learning approach confirms that prior beliefs about the wage impact of the Mariel boatlift are most relevant in shaping responses to the treatment.

[26] We specify the bin size of the final nodes to be at least 200; and use 50 cross-validations sets for the out-of-sample predictions, i.e. we randomly partition the data into equally-sized sub-samples. One of the folds is used for evaluation, while the remaining folds are used for fitting the tree.

Figure 1.5: Heterogeneous Treatment Effects: Machine-learning Classification and Regression Trees



Notes: This figure displays causal trees using Classification and Regression Trees (CART) (Athey and Imbens, 2016). "Wage prior (LS)" measures people's beliefs about the impact of the Mariel boatlift on wages of low-skilled workers in Miami. The responses are on a 5 point-scale ranging from (1) strongly decrease to (5) strongly increase. "Confidence in Wage prior (LS)" measures people's confidence in their prior beliefs about the wage effects of the Mariel boatlift. Responses are measured on a five point scale ranging from (1) very unsure to (5) very sure. Confidence in Unemployment Prior (HS) measure people's confidence in their prior beliefs about the employment effects of the Mariel boatlift. Responses are measured on a five point scale ranging from (1) very unsure to (5) very sure. White takes value one for respondents identifying as Whites, and zero otherwise. "HH size" denotes people's household size. Not unemployed takes value 1 of respondents self-report not to be unemployed. Northeast takes value 1 if respondents report to have their current residence in the Northeast, and zero otherwise. West takes value 1 if respondents report to have their current residence in the West, and zero otherwise. Log income denotes the log of respondents' self-reported household income in US dollars (log income of 11 corresponds to an annual household income of approximately 60,000 USD). The numbers at the final nodes denote the z-scored treatment effects for the different subgroups defined by the causal tree. We include all variables displayed in the balance Table A.16 in the machine learning algorithm, and we set the minimum number of the final node as 200 and the number of cross validation sets as 50.

1.4 Discussion

Section 1.3 established that labour market concerns have a causal effect on support for immigration. This section extends this analysis by exploring two additional questions: (i) whether labour market concerns operate through self-interested concerns or sociotropic concerns about the nationwide impact of immigration, and (ii) whether economic or cultural concerns are more important in the formation of immigration preferences. We end the section by discussing potential reasons as to why our results differ from the previous literature.

1.4.1 Self-interested vs. sociotropic labour market concerns

To test whether self-interested or sociotropic labour market concerns are more important in the formation of immigration attitudes, we use an OLS regression with control group respondents, where we simultaneously include beliefs about the labour market impact of immigration on both the respondent's own household as well as on "most Americans".

Beliefs about the labour market impact on most Americans are stronger predictors of support for immigration than beliefs about the labour market impact on people's own household (see Table A.11). While a one standard deviation change in beliefs about the wage impact of immigration on most Americans is associated with a 0.12 to 0.25 standard deviation change in attitudes towards different types of immigrants, the corresponding point estimates for beliefs about the wage impact on people's own household are either zero or negative. The same pattern emerges when we control for be-

liefs about the cultural impact of immigration.[27] Overall, these estimates suggest an important role for sociotropic labour market concerns in the formation of immigration attitudes.

1.4.2 Economic vs. cultural concerns

To test whether economic or cultural concerns are more important in forming attitudes towards immigration, we use regressions with control group respondents, simultaneously including beliefs about the overall economic impact of immigration ("do you think the overall effect would be positive or negative for the finances of most Americans?") and the overall cultural impact of immigration ("how do you think that increasing the number of low-skilled immigrants to the United States would affect American culture and society as a whole?").

Table A.12 shows that the standardised coefficients on economic concerns are 2–4 times larger than the analogous coefficients for cultural concerns. For instance, while a one standard deviation change in beliefs about the overall economic impact of low-skilled immigration is associated with a 0.59 standard deviation change in support for low-skilled immigration, the corresponding point estimate for a one standard deviation change in beliefs about the cultural impact of low-skilled immigration is 0.16.

These findings contrast those of Card et al. (2012) who found that, in European countries, cultural concerns were 2–5 times stronger predictors of attitudes towards immigration than economic concerns. There are several reasons that could explain the difference in findings. First, the difference

[27] We also replicate these correlational patterns using data from a probability-based sample of the US population (the 2014 wave of the General Social Survey; see Figure A.5).

could reflect different compositions of immigrants in the US and Europe. Second, the difference could reflect that we use different proxies for cultural concerns. While we ask our respondents how they think immigration would affect American culture and society as a whole, Card et al. (2012) ask specific questions about religion, language, social tension, customs, and traditions.

As a robustness check, we replicate the results using data from the 2014 wave of the General Social Survey (GSS), a probability-based survey which is designed to be fully representative of the US population.[28] Economic concerns are about a seven times stronger predictor of attitudes towards immigration than cultural concerns (see Figure A.5).

1.4.3 Why do our results differ from the current consensus?

The current consensus, as formulated by Hainmueller and Hopkins (2014b, p. 241), is that labour market concerns are not a significant driver of mass attitudes towards immigration. In the following, we review some of the main results from seminal studies in this literature (Hainmueller and Hiscox, 2010; Hainmueller et al., 2015; Hainmueller and Hopkins, 2014a) and discuss why our results differ.

To investigate whether labour market concerns drive support for immigration, Hainmueller and Hiscox (2010) test whether people's education levels predict people's support for immigrants of different skill levels. Based on

[28] To reduce measurement error, we use two questions to create a standardised index of cultural concerns: (i) "American culture is generally undermined by immigrants" and (ii) "Immigrants make America more open to new ideas and cultures". We create a similar index of economic concerns based on the following two questions: "Immigrants take jobs away from people who were born in America" and (ii) "Immigrants are generally good for America's economy".

their main finding that highly educated respondents are more supportive of immigration of both high-skilled and low-skilled immigrants than uneducated respondents, Hainmueller and Hiscox (2010) conclude that self-interested labour market concerns are not an important driver of people's attitudes towards immigration. Hainmueller et al. (2015) draw similar conclusions using proxies for skill levels that are less crude than educational attainment.[29]

For this to be a valid test of the relevance of labour market concerns, two assumptions must hold. First, people's beliefs should be in line with factor proportion models which predict that people gain (lose) from immigration of workers with different (similar) skill levels from their own. We find mixed evidence on the validity of this assumption.[30] Second, labour market concerns should mainly operate through self-interested concerns. As shown in Section 1.4.1, our estimates suggest that labour market concerns mainly operate through sociotropic concerns about how immigration affects the labour market outcomes of most Americans.

The conclusions in Hainmueller and Hiscox (2010) and Hainmueller et al. (2015) are based on correlations between skill levels and support for immigration. To address causality, Hainmueller and Hopkins (2014a) use conjoint experiments where they measure support for hypothetical immigrants with randomly assigned characteristics, such as their education levels and whether they plan to find work. The main finding in Hainmueller and Hopkins (2014a) is that Americans' preferences for which immigrants to admit

[29]We replicate this result using our data with both objective and subjective measures of skill level (see Table A.9).

[30]Table A.10 tests this assumption by regressing people's beliefs about how their own household is affected by low-skilled and high-skilled immigration on their educational attainment and their subjective skill levels. These regressions offer some support for this assumption, but the correlations are weak and not always significant.

vary little with their own education and labour market position. The authors interpret these findings as reflecting a limited role of labour market concerns in driving support for immigration. Our data on subjective beliefs demonstrate the difficulty of separating such competing explanations without exogenous variation in beliefs. For instance, our respondents think that high-skilled immigrants have a 0.32 of a standard deviation more positive impact on American values and traditions than low-skilled immigrants. The claim that views on immigration are primarily rooted in cultural concerns (Hainmueller et al., 2015; Sniderman et al., 2004) fails to acknowledge that that people's economic perceptions depend on the cultural characteristics of immigrants.

1.5 Concluding remarks

In this paper, we present evidence that labour market concerns causally affect public support for immigration. Signatures on real online petitions demonstrate that the research evidence affects real political behavior. We find that a one standard deviation change in beliefs about the economic impact of immigration changes attitudes towards increasing low-skilled immigration by 0.61 of a standard deviation. The effects also persist in an obfuscated follow-up study where concerns about differential demand effects across the treatment and control group are mitigated. These findings challenge the current consensus that labour market concerns are not a quantitatively important determinant of mass attitudes towards immigration.

Our evidence complements a literature documenting how immigration affects electoral support for anti-immigrant parties (Becker and Fetzer, 2016; Dehdari, 2017; Dustmann et al., 2018; Halla et al., 2017). There has been a

big debate about whether recent opposition to immigration in the US and Europe is due to the perceived economic or cultural threats of immigration. While natural experiments studying the impact of immigration on voting outcomes cannot identify the underlying motivations of the voters, our findings suggest that labour market concerns may be an important mechanism at play.

We believe that there are several promising pathways for future research on the role of labour market concerns in driving support for immigration. First, it would be interesting to examine whether anecdotal evidence on the effect of the Mariel boatlift would have similar effects on support for immigration. Second, it would be interesting to examine whether exogenously shifting only people's beliefs about the effect of immigrants on one's own household would confirm our correlational results which suggest that labour market concerns primarily operate through sociotropic considerations about how immigration affects most people's labour market outcomes.

Appendices of Chapter 1

Summary of the Appendices

Section 1.A provides a conceptual framework. Section 1.B provides additional descriptives on our respondents beliefs. Figure A.1 displays the distribution of preferences and beliefs using control group respondents. Figure A.2 displays the correlates of demographics with people's pre-treatment beliefs about the wage and employment impact of the Mariel boatlift. Figure A.3 shows treatment effects in absolute and relative changes. Figure A.4 shows heterogeneous updating of beliefs about the labour market effects of immigrants. Finally, Figure A.5 displays correlations on labour market concerns and cultural concerns using the General Social Survey.

Section 3.B provides further evidence on the mechanisms, and descriptives. Table A.1 summarizes treatment effects on people's beliefs about how their own household's and most American's wages, employment, tax burden and American culture are affected by low-skilled immigration. Table A.2 summarizes treatment effects on people's beliefs about how their own household's and most American's wages, employment, tax burden and American culture are affected by low-skilled immigration for the subsample of respondents that are currently in the labour force. Table A.3 shows the main results on policy preferences restricting the sample to those who also completed the follow-up survey. Table A.4 shows results on the IV regressions

instrumenting beliefs about the labour market impact of immigration with the treatment assignment using the full sample of respondents. Table A.5 shows results on the IV regressions instrumenting beliefs about the overall economic impact of immigration with the treatment assignment using the full sample of respondents.

Table A.6 displays the gap in beliefs about the economic effects of immigration between Republicans and Democrats. Table A.7 describes correlates of people's beliefs about the labour market impact of the Mariel boatlift. Table A.8 shows correlates of people's attitudes. Table A.9 examines how skill level correlates with attitudes towards immigrants. Table A.10 investigates whether people's beliefs are in line with the factor-proportion model. Table A.11 examines the relative importance of self-interested vs. group-level concerns in shaping immigration preferences. Table A.12 explores the relative importance of economic vs. cultural concerns in shaping immigration preferences. Table A.13 shows effects of the demand treatment on immigration attitudes.

Section 1.D provides evidence on covariate balance and describes the sample. Table A.15 provides descriptive statistics. Table A.16 shows covariate balance for the main sample, while Table A.17 provides evidence on covariate balance for the sample in the follow-up study. Table A.18 examines correlates of attrition in the follow-up study.

Section 1.D.1 examines heterogeneity in treatment effects. Table A.19 explores heterogeneity for our main self-reported outcome questions by people's self-perceived skill level, their political affiliation, and their prior belief about the effect of the Mariel boatlift. Table A.20 examines heterogeneous treatment effects along these dimensions on people's intention to sign the

petition, while Table A.21 examines heterogeneity using the main outcome questions in the follow-up study. Table A.23 and Table A.22 explore heterogeneous treatment effects on people's beliefs about the labour market impact of immigration. Table A.24 shows results on people's beliefs about the overall economic impact of immigration.

In Section 1.E, Figure A.6 describes our respondents' beliefs about the survey and beliefs about the relevance, accuracy, and trustworthiness of the research evidence. Table A.25 explores how beliefs about the research study are correlated with support for immigration. Table A.26 examines how standard demographics are correlated with beliefs about the research study.

Finally, Section 1.F provides screenshots of the online petition, the invitation email for the follow-up study, and the consent forms in the main study and the obfuscated follow-up study.

1.A Conceptual framework

We here present a simple theoretical framework, adapted from Card et al. (2012), on how people report their policy views on immigration. Individual i's indirect utility function depends on her net income and the characteristics of her community:

$$u_i(w_i - t_i, \mathbf{c}_i)$$

where w_i is i's gross income; t_i is taxes; and $\mathbf{c}_i$ is a vector of characteristics of i's community, such as ethnic composition and religious diversity.[31] Throughout the paper, we refer to net income, $w_i - t_i$, and community characteristics, $\mathbf{c}_i$, as respectively "economic" and "cultural" factors. When

[31] Card et al. (2012) use the term "compositional amenities" for these factors.

asked about their self-reported policy preferences on immigration, individuals compare a hypothetical scenario with more immigrants (w_i',t_i',$\mathbf{c}_i'$) to the current situation (w_i°,t_i°,$\mathbf{c}_i^\circ$). The individuals then report a transformation of the gap in the indirect utilities under the different scenarios:

$$y_i = g_i[(w_i', t_i', \mathbf{c}_i') - (w_i^\circ, t_i^\circ, \mathbf{c}_i^\circ)]$$

where g_i is an individual-specific (and increasing) response function. Taking a first-order approximation, the observed response of individual i is

$$y_i = g_i' \times \frac{\partial u_i}{\partial w_i}(\Delta w_i - \Delta t_i) + g_i' \times \frac{\partial u_i}{\partial a_i}\Delta a_i + \varepsilon_i \tag{1.1}$$

where $\Delta w_i = w_i' - w_i^\circ$ is the difference in perceived gross earnings between the alternative scenarios (with analogous definitions for Δt_i and $\Delta \mathbf{c}_i$) and ε_i is an idiosyncratic error term. In this model, we can define labour market concerns as $g_i' \times \frac{\partial u_i}{\partial w_i}\Delta w_i$.

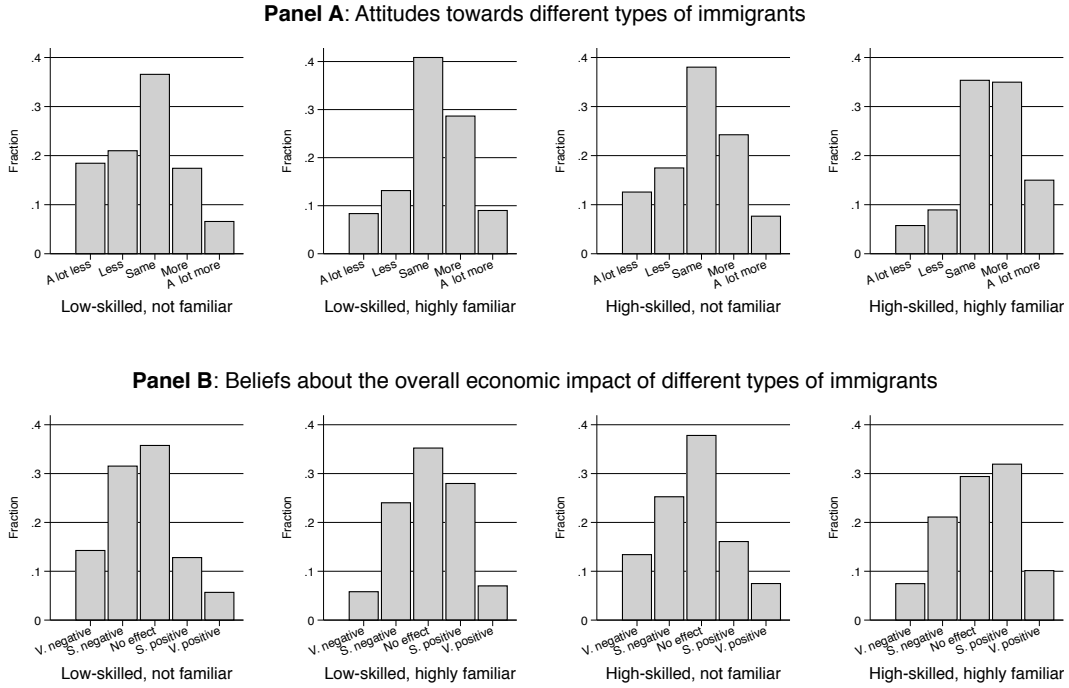
To identify the role of labour market concerns in driving support for immigration without making any restrictive assumptions, one needs an instrument that exogenously shifts people's perceived labour market impact of immigration, Δw. We use a randomised information treatment to achieve this. Our treatment informs people that immigration has virtually no impact on the labour market; i.e, that $\Delta w \approx 0$. The model thus predicts that the treatment should increase support for immigration if people on average expect negative labour market impacts of immigration and treated respondents (at least partly) trust the information.

The treatment needs to satisfy two conditions to be a valid instrument for labour market concerns. First, the treatment should change beliefs about

the labour market impact immigration, Δw_i, but not simultaneously change beliefs about the impact of increased immigration on taxes, Δt_i, or culture, $\Delta \mathbf{c}_i$. We ask post-treatment questions about the expected impact of immigration on both taxes and culture to test these assumptions. Second, the treatment should not affect the reporting function, g_i. While this assumption is difficult to test, we argue that concerns about temporary changes in the reporting function are mitigated through the obfuscated follow-up.

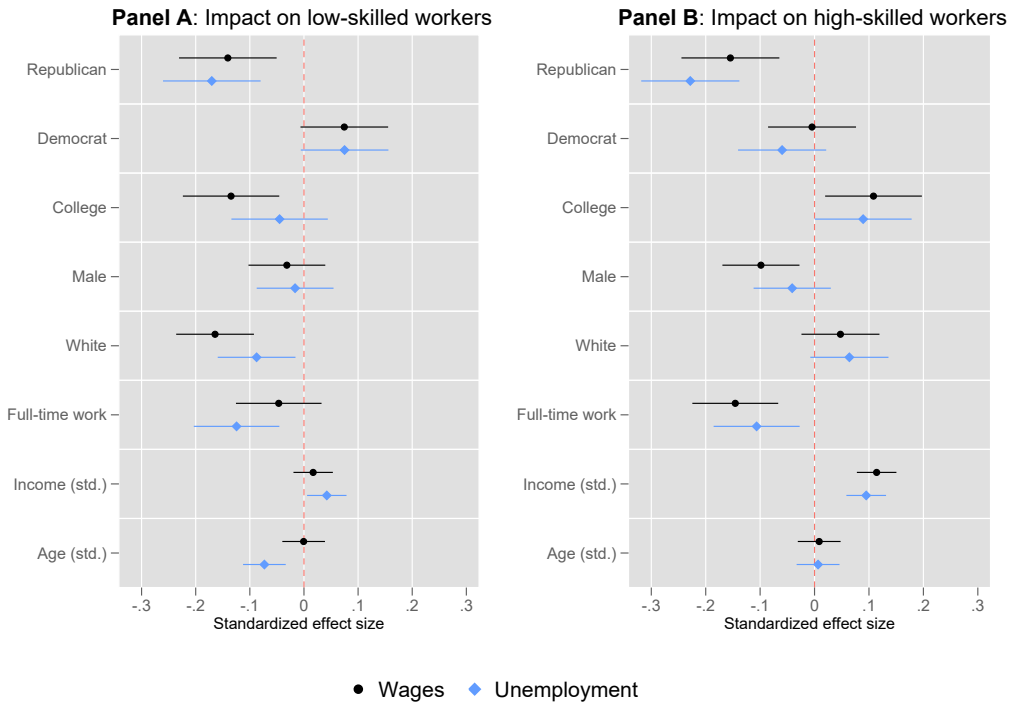
1.B Additional Figures

Figure A.1: Heterogeneity in policy views and beliefs about immigration



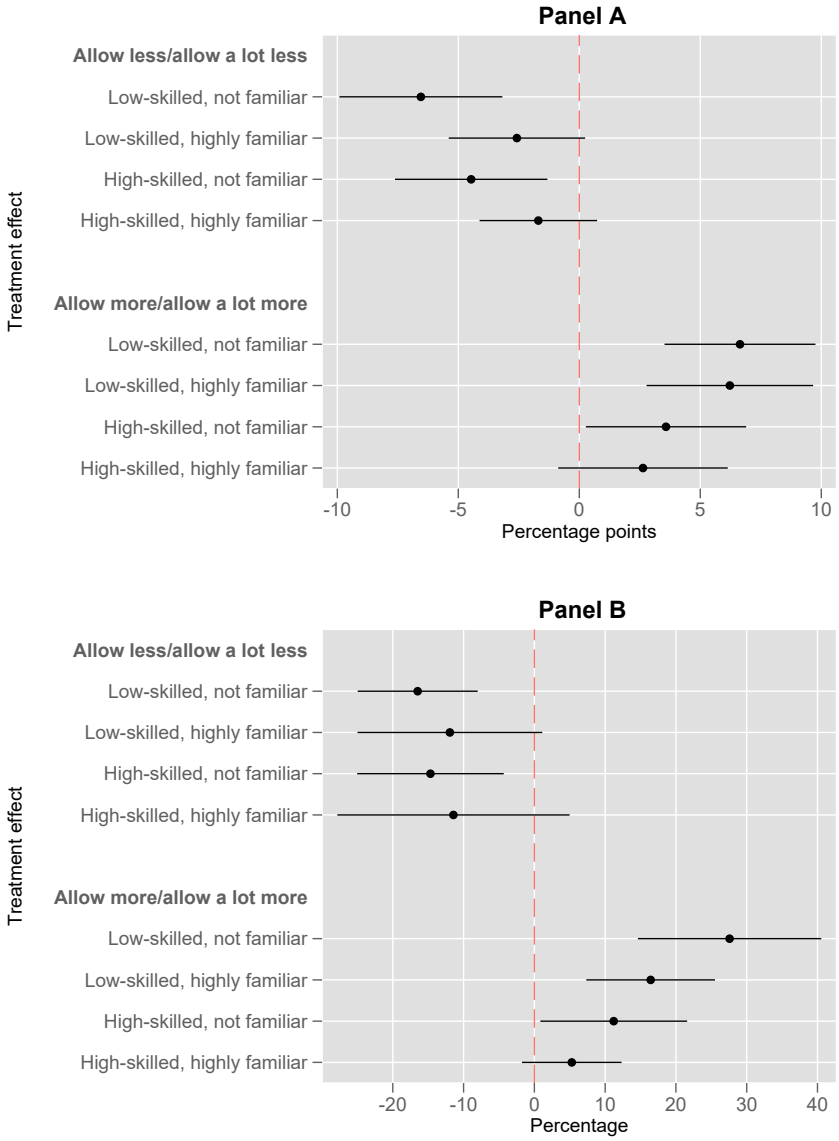
Notes: **Panel A** shows the distribution of policy views on whether to increase or decrease the number of low-skilled/high-skilled immigrants that are highly familiar/not familiar with American values and traditions. **Panel B** shows the distribution of beliefs about the "overall economic impact on most Americans" for the different types of immigrants. Both panels only include responses from respondents in the control group.

Figure A.2: Correlates of pre-treatment beliefs about the Mariel boatlift



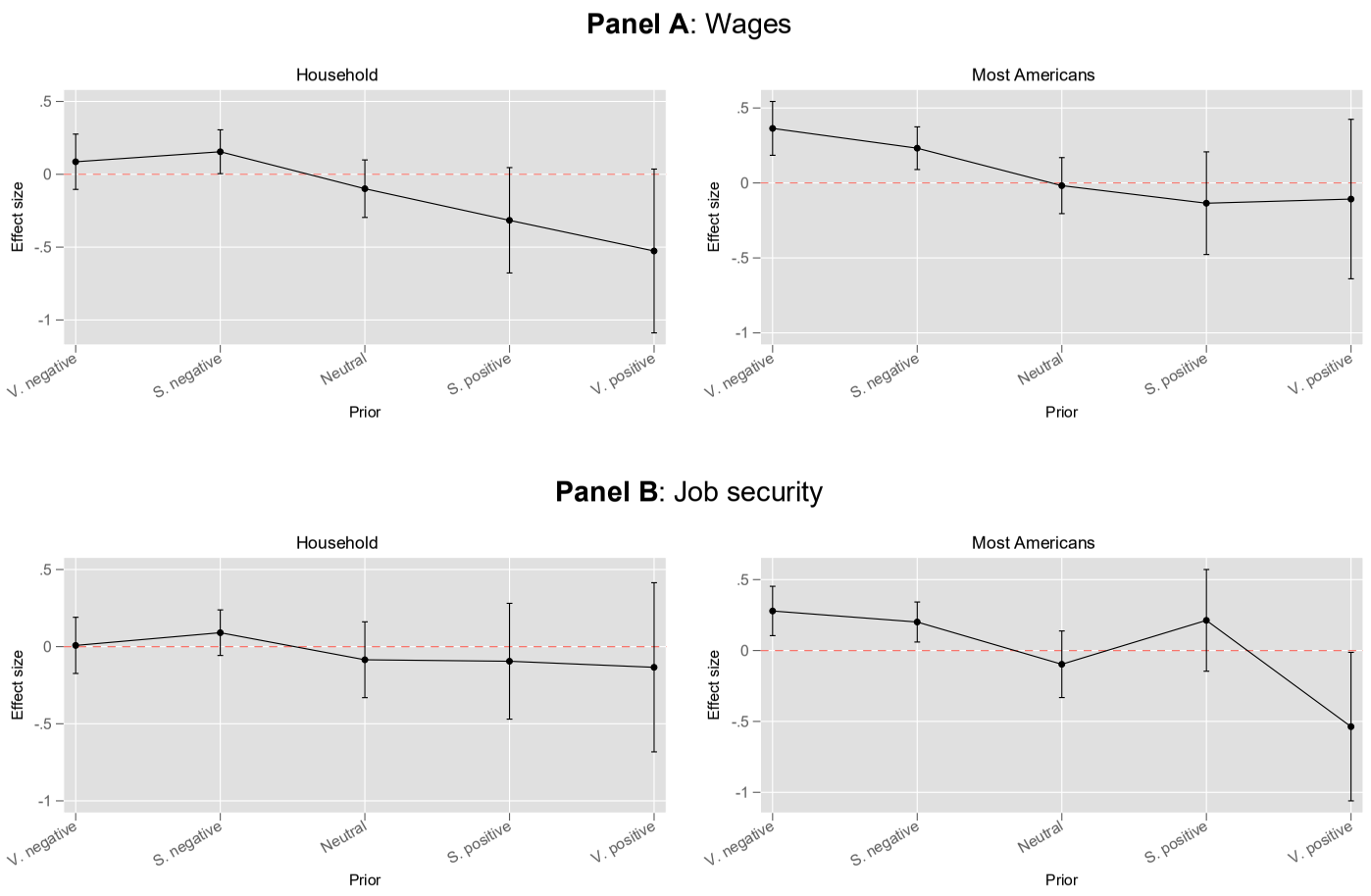
Notes: The dots indicate the mean values of the estimated multiple regression coefficients. The lines indicate 95 percent confidence interval of the mean. In **Panel A**, the outcome variables are people's beliefs about the labour market impact of the Mariel boatlift on low-skilled workers; in **Panel B**, the outcome variables are beliefs about the labour market impact on high-skilled workers.

Figure A.3: Treatment effects on attitudes towards immigrants



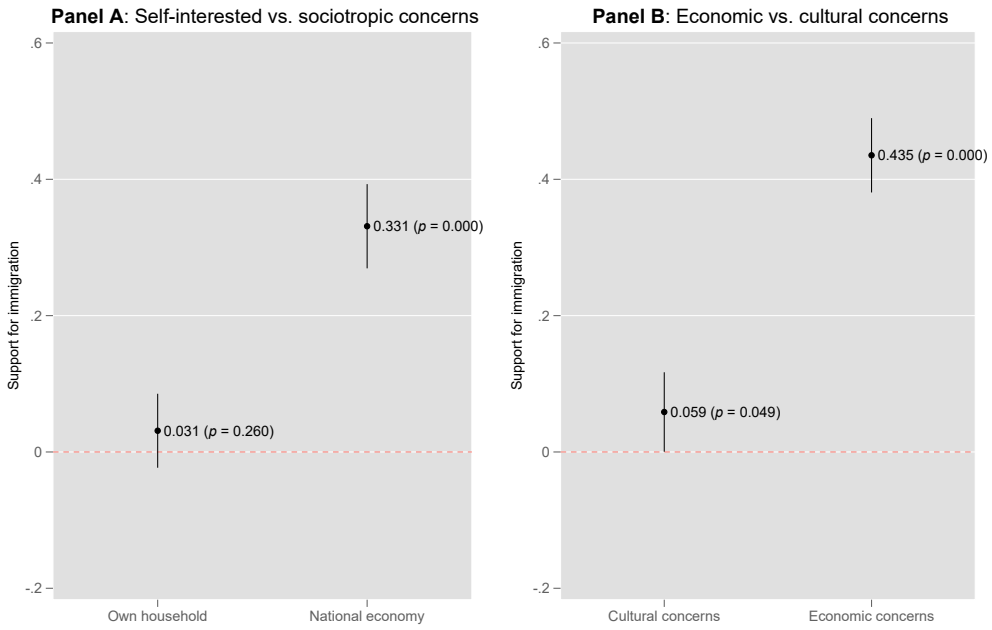
Notes: The figure shows treatment effects on the fraction of respondents who answer that they want to allow more/allow a lot more and allow less/a lot less of low-skilled/high-skilled immigrants that are highly familiar/not familiar with American values and traditions. **Panel A** shows the treatment effects in percentage points changes. **Panel B** shows the treatment effects in percent changes. Lines indicate 95 percent confidence intervals.

Figure A.4: Heterogeneity in belief updating



Notes: The figure shows standardised treatment effects on beliefs about the wage and employment effects immigration by pre-treatment beliefs about the wage impact of immigration. The outcomes are based on the follow questions: how do you think admitting more low-skilled immigrants would affect (i) "wages," and (ii) "job opportunities and job security," for their own household as well as for most Americans. 95 percent confidence intervals are indicated.

Figure A.5: Self-interested vs. group level concerns: General Social Survey



Notes: Panel A uses data from the 1994 wave of the General Social Survey. Self-interested labour market concerns are measured with people's answer to the following question: "What about immigrants? Is it very likely, somewhat likely, somewhat unlikely, or very unlikely that you or anyone in your family won't get a job or promotion while an equally or less qualified immigrant employee receives one instead?" We proxy for people's group-level labour market concerns with people's response to the following question: "How much do you agree or disagree with each of the following statements? Immigrants take jobs away from people who were born in America". Panel B uses data from the 2014 wave of the General Social Survey. *Support for immigration* is the standardised response as to whether "the number of immigrants to America nowadays" should be "decreased a lot" to "increased a lot". *Economic concerns* is a standardised index defined by people's agreement to the following two statements: (i) "Immigrants take jobs away from people who were born in America" and (ii) "Immigrants are generally good for America's economy" (higher values of the index imply more more optimistic beliefs). Cultural concerns is defined by people's agreement to the following two statements: (i) "American culture is generally undermined by immigrants" and (ii) "Immigrants make America more open to new ideas and cultures". All variables have been z-scored and higher values imply more optimistic beliefs. Lines indicate 95 percent confidence intervals.

1.C Additional Tables

Table A.1: Beliefs about the effects of low-skilled and high-skilled (post-treatment)

	Wages		Employment		Fiscal burden		Culture
	Own household	Most Americans	Own household	Most Americans	Own household	Most Americans	
Panel A: Effect of low-skilled							
Treatment	0.037	0.178***	0.019	0.148***	0.027	0.029	0.033
	(0.048)	(0.046)	(0.049)	(0.047)	(0.049)	(0.047)	(0.049)
Observations	1489	1489	1484	1484	1479	1479	1469
Panel B: Effect of high-skilled							
Treatment	0.064	0.185***	0.155***	0.258***	0.054	0.075	0.045
	(0.047)	(0.048)	(0.048)	(0.048)	(0.047)	(0.047)	(0.048)
Observations	1487	1487	1481	1481	1475	1475	1461

Notes: The table shows OLS regression results where the dependent variables are beliefs about the economic and cultural impact of different types of immigrants. Respondents were asked how they thought admitting more low-skilled/high-skilled immigrants would affect (i) "wages," (ii) "job opportunities and job security," and (iii) "taxes" for their own household as well as for most Americans. They were also asked about how they thought admitting more low-skilled/high-skilled immigrants "would affect American culture and society as a whole". We randomised whether respondents answered these questions for low-skilled or high-skilled immigrants. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views on the effect of immigration. The outcome variables are z-scored using the mean and standard deviation in the control group. We use the same controls as described in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.2: Mechanisms: Effects of low-skilled and high-skilled: Sample of people in the labour force

	Wages		Employment		Fiscal burden		Culture
	Own household	Most Americans	Own household	Most Americans	Own household	Most Americans	
Panel A: Effect of low-skilled							
Treatment	0.035	0.156***	0.022	0.117**	0.022	-0.000	0.033
	(0.061)	(0.056)	(0.061)	(0.057)	(0.060)	(0.057)	(0.059)
Observations	1070	1070	1065	1065	1060	1060	1050
Panel B: Effect of high-skilled							
Treatment	0.074	0.173***	0.173***	0.229***	0.064	0.074	0.035
	(0.060)	(0.058)	(0.061)	(0.058)	(0.057)	(0.057)	(0.058)
Observations	1083	1083	1077	1077	1071	1071	1057

Notes: The table shows OLS regression results for the subsample of our respondents that are in the labour force. The dependent variables are beliefs about the economic impact of low-skilled and high-skilled immigrants. Respondents were asked how they thought admitting more low-skilled/high-skilled immigrants would affect (i) "wages," (ii) "job opportunities and job security," and (iii) "taxes" for their own household as well as for most Americans. We randomised whether respondents answered these questions with respect to low-skilled or high-skilled immigrants. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views regarding the effect of immigration. The outcome variables are z-scored using the mean and standard deviation in the control group. All regressions include controls as described in the pre-analysis plan and Table 1.1, but the results are not sensitive to the inclusion of controls. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.3: Self-reported attitudes towards immigration (post-treatment) – sample of respondents who completed the follow-up

	Low-skilled (Main Study)			High-skilled (Main Study)			Follow-up	
	Not familiar	Familiar	Index	Not familiar	Familiar	Index	Low-skilled	High-skilled
Panel A: With controls								
Treatment	0.144***	0.106**	0.125***	0.089**	0.045	0.067*	0.095**	0.059
	(0.040)	(0.042)	(0.037)	(0.040)	(0.042)	(0.037)	(0.040)	(0.042)
Adjusted p-value	[0.001]	[0.005]		[0.058]	[0.163]			
Observations	2075	2075	2075	2075	2075	2075	2075	2075
Panel B: Without controls								
Treatment	0.153***	0.119***	0.136***	0.103**	0.051	0.077**	0.103**	0.062
	(0.044)	(0.044)	(0.040)	(0.043)	(0.043)	(0.039)	(0.043)	(0.043)
Adjusted p-value	[0.001]	[0.003]		[0.035]	[0.132]			
Observations	2075	2075	2075	2075	2075	2075	2075	2075

Notes: The table shows OLS regression results where the dependent variables are attitudes to the different types of immigrants and uses respondents who completed both the initial survey and the follow-up survey. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. The indices are created by taking the mean of the responses to immigrants with different familiarity with American values and traditions for each skill level. Adjusted p-values are in brackets. "Treatment" is an indicator equal to 1 if respondents received the research evidence. Controls include gender, age, ethnicity, region, household size, household income, education, employment status, party affiliation, whether the respondent was born in the US, whether the subject's parents were born in the US, self-perceived skill-level, and pre-treatment beliefs about the labour market impact of low-skilled (or high-skilled) immigration and are coded as described in the pre-analysis plan. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.4: The causal effect of labour market concerns: Full sample

	Low-skilled			High-skilled		
	Not familiar	Familiar	Index	Not familiar	Familiar	Index
Panel A: IV						
Labor Market Concerns	0.967***	0.620**	0.794***	0.556***	0.165	0.361*
	(0.313)	(0.305)	(0.280)	(0.213)	(0.222)	(0.192)
Adjusted p-value	[0.006]	[0.014]	[0.006]	[0.027]	[0.179]	[0.064]
Observations	1472	1472	1472	1466	1468	1468
Panel B: IV with additional controls						
Labor Market Concerns	1.032***	0.701*	0.867***	0.588*	-0.033	0.278
	(0.358)	(0.358)	(0.323)	(0.329)	(0.337)	(0.291)
Adjusted p-value	[0.011]	[0.016]	[0.011]	[0.282]	[1.000]	[0.513]
Observations	1472	1472	1472	1466	1468	1468
Panel C: OLS						
Labor Market Concerns	0.383***	0.308***	0.308***	0.355***	0.339***	0.339***
	(0.028)	(0.030)	(0.030)	(0.029)	(0.030)	(0.030)
Adjusted p-value	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Observations	1472	1472	1472	1466	1468	1468
Panel D: OLS with additional controls						
Labor Market Concerns	0.302***	0.266***	0.308***	0.208***	0.170***	0.170***
	(0.035)	(0.039)	(0.030)	(0.041)	(0.043)	(0.043)
Adjusted p-value	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Observations	1472	1472	1472	1466	1468	1468

Notes: These regressions are based on the subsample of respondents who thought that the Mariel boatlift had either negative or no effects on wages and unemployment. The dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". "Economic Concerns" is given by the answers to the following question: "When you think about all of the potential positive and negative economic effects of increasing the number of high-skilled/low-skilled immigrants that are not familiar/highly familiar with American values and traditions coming to the United States, do you think the overall effect would be positive or negative for the finances of most Americans?" The question was answered on a five point scale from 1: "The overall economic effect would be very negative for most Americans" to 5: "The overall economic effect would be very positive for most Americans". In Panel A, we instrument "Economic Concerns" with an indicator for whether our respondent received the research evidence. In Panel B, we run OLS regressions. We randomised whether people were asked about the impact of low-skilled or high-skilled immigrants. We use the same controls as described in Table 1.1. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses.

Table A.5: The causal effect of economic concerns

	Low-skilled		High-skilled	
	Not familiar	Familiar	Not familiar	Familiar
Panel A: IV				
Economic Concerns	0.717***	0.617**	0.813***	0.264
	(0.184)	(0.255)	(0.298)	(0.346)
Adjusted p-value	[0.001]	[0.007]	[0.012]	[0.286]
Observations	1472	1472	1466	1468
Panel B: OLS				
Economic Concerns	0.526***	0.446***	0.493***	0.465***
	(0.022)	(0.023)	(0.023)	(0.024)
Adjusted p-value	[0.001]	[0.001]	[0.001]	[0.001]
Observations	1472	1472	1466	1468

Notes: The dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". "Economic Concerns" is given by the answers to the following question: "When you think about all of the potential positive and negative economic effects of increasing the number of high-skilled/low-skilled immigrants that are not familiar/highly familiar with American values and traditions coming to the United States, do you think the overall effect would be positive or negative for the finances of most Americans?" The question was answered on a five point scale from 1: "The overall economic effect would be very negative for most Americans" to 5: "The overall economic effect would be very positive for most Americans". In Panel A we instrument "Economic Concerns" with an indicator for whether our respondent received the research evidence. In Panel B we run OLS regressions. We randomised whether people were asked about the impact of low-skilled or high-skilled immigrants. We use the same controls as described in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.6: Beliefs about the impact of immigrants and the Republican–Democrat Gap

	Low-skilled Index			High-skilled Index		
Republican	-0.678***	-0.332***	-0.297***	-0.574***	-0.283***	-0.194**
	(0.081)	(0.072)	(0.072)	(0.088)	(0.078)	(0.079)
Beliefs: overall economic impact		0.563***	0.545***		0.467***	0.385***
		(0.038)	(0.042)		(0.044)	(0.050)
Beliefs: overall cultural impact			0.071*			0.205***
			(0.038)			(0.047)

Notes: This table shows regression results using respondents from the control group who are not Independents. In columns (1) to (3) the outcome variable is an index of support for low-skilled immigration. In columns (4) to (6) the outcome variable is an index of support for high-skilled immigration. Republican takes value 1 for respondents who self-identify as Republican. "Beliefs: overall economic impact" is a z-scored variable on a 5-point scale and takes higher values for respondents with more optimistic beliefs about the economic impact of immigration. "Beliefs: overall cultural impact" is a z-scored variable on a 5-point scale and takes higher values for respondents with more optimistic beliefs about the cultural impact of immigration.

Table A.7: Correlates of beliefs about the impacts of the Mariel boatlift

	Effects on Low-skilled		Effect on High-skilled	
	Wages	Unemployment	Wages	Unemployment
Republican	-0.218***	-0.227***	-0.125***	-0.158***
	(0.046)	(0.046)	(0.047)	(0.047)
Independent	-0.070*	-0.050	0.012	0.052
	(0.042)	(0.042)	(0.041)	(0.041)
Age	-0.000	-0.004***	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Log(Income)	-0.016	0.011	0.142***	0.119***
	(0.025)	(0.025)	(0.025)	(0.026)
High-skilled	-0.019	-0.057	-0.178***	-0.146***
	(0.049)	(0.048)	(0.048)	(0.047)
Employed Full-Time	-0.035	-0.085*	-0.173***	-0.131***
	(0.052)	(0.051)	(0.051)	(0.051)
Employed Part-Time	0.065	0.158**	0.024	-0.011
	(0.072)	(0.079)	(0.072)	(0.068)
Unemployed	0.002	0.104	0.031	-0.103
	(0.107)	(0.101)	(0.109)	(0.102)
Self-Employed	-0.007	-0.003	-0.179**	-0.081
	(0.083)	(0.090)	(0.084)	(0.082)
Student	0.047	-0.050	0.171	0.261*
	(0.125)	(0.125)	(0.138)	(0.143)
High Education	-0.261***	-0.116	-0.044	-0.035
	(0.073)	(0.072)	(0.075)	(0.074)
Male	-0.044	-0.022	-0.078**	-0.023
	(0.037)	(0.037)	(0.037)	(0.037)
Share of immigrants zip code	0.444***	0.266**	-0.066	-0.010
	(0.143)	(0.135)	(0.138)	(0.135)
Observations	3130	3130	3130	3130

Notes: The outcome variables are people's pre-treatment beliefs about the effect of the Mariel boatlift on wages and unemployment of low-skilled incumbent workers and high-skilled incumbent workers. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views regarding the effect of immigration. The outcome variables are z-scored using the mean and standard deviation in the control group. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.8: Correlates of attitudes towards immigrants

	Low-skilled		High-skilled	
	Not familiar	Familiar	Not familiar	Familiar
Republican	-0.602***	-0.696***	-0.460***	-0.731***
	(0.067)	(0.062)	(0.067)	(0.063)
Independent	-0.230***	-0.330***	-0.182***	-0.318***
	(0.055)	(0.055)	(0.056)	(0.055)
Age	-0.008***	-0.010***	-0.008***	-0.010***
	(0.002)	(0.002)	(0.002)	(0.002)
Log(Income)	0.063*	-0.013	0.066*	0.048
	(0.035)	(0.034)	(0.035)	(0.033)
High-skilled	0.150**	0.177***	0.106	0.097
	(0.065)	(0.064)	(0.065)	(0.063)
Employed Full-Time	0.036	0.053	-0.069	-0.016
	(0.073)	(0.069)	(0.074)	(0.068)
Employed Part-Time	0.058	0.147	-0.041	0.029
	(0.107)	(0.102)	(0.104)	(0.100)
Unemployed	0.030	0.186	-0.080	0.090
	(0.144)	(0.129)	(0.142)	(0.140)
Self-Employed	-0.129	-0.080	-0.218*	-0.016
	(0.114)	(0.102)	(0.118)	(0.106)
Student	-0.039	0.023	0.046	0.200
	(0.156)	(0.165)	(0.145)	(0.161)
High Education	-0.011	-0.056	-0.003	-0.051
	(0.097)	(0.093)	(0.097)	(0.093)
Male	-0.042	-0.017	0.121**	0.084*
	(0.051)	(0.049)	(0.051)	(0.049)
Share of immigrants zip code	-0.161	0.017	-0.221	-0.017
	(0.175)	(0.167)	(0.173)	(0.168)
Observations	1571	1571	1571	1571

Notes: This table uses data from respondents in the control group. The table shows OLS regression results where the dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.9: Stylized facts: Attitudes towards immigrants by self-perceived skill level and objective skill-level

	Low-skilled		High-skilled	
	Not familiar	Familiar	Not familiar	Familiar
Panel A:				
High skilled	0.187***	0.185***	0.121**	0.176***
	(0.068)	(0.060)	(0.060)	(0.062)
Panel B:				
Higher Education	0.001	0.085***	0.046	0.099***
	(0.035)	(0.031)	(0.031)	(0.032)
Observations	1467	1467	1467	1467

Notes: This table uses data from respondents in the control group. The table shows OLS regression results where the dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. "High-skilled" takes value one for respondents who self-categorize as high-skilled, and zero otherwise. "Higher Education" takes value 1 for people who dropped out of high-school, 2 for people who completed at most high school, 3 for people who completed at least some college and 4 for people who at least finished a bachelor's degree. We use the same controls as described in Table 1.2. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.10: Stylized facts: Beliefs about the impact of high and low-skilled immigrants on own household by self-perceived skill level and objective skill-level

	Low-skilled immigrants		High-skilled Immigrants	
	own wages	own employment	own wages	own employment
Panel A:				
High skilled	0.016	0.202**	-0.171**	-0.106
	(0.086)	(0.085)	(0.084)	(0.084)
Panel B:				
Higher Education	0.144***	0.172***	0.004	-0.043
	(0.042)	(0.042)	(0.045)	(0.045)
Observations	756	756	711	711

Notes: This table uses data from respondents in the control group. The dependent variables are beliefs about the economic impact of low-skilled and high-skilled immigrants on people's own household. Respondents were asked how they thought admitting more low-skilled/high-skilled immigrants would affect (i) "wages," (ii) "job opportunities and job security," and (iii) "taxes" of their own household. We randomised whether respondents answered these questions with respect to low-skilled or high-skilled immigrants. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views regarding the effect of immigration. The outcome variables are z-scored using the mean and standard deviation in the control group. All regressions include controls as described in the pre-analysis plan and Table 1.1, but the results are not sensitive to the inclusion of controls. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.11: Self-interested labour market concerns vs. group-level labour market concerns

	Low-skilled		High-skilled		Low-skilled		High-skilled	
	Not familiar	Familiar	Not familiar	Familiar	Not familiar	Familiar	Not familiar	Familiar
Effect of immigrants on own household's wage	-0.079	0.006	-0.075	-0.007	-0.093	-0.000	-0.088	-0.022
	(0.066)	(0.068)	(0.061)	(0.063)	(0.065)	(0.067)	(0.059)	(0.061)
Effect of immigrants on most Americans' wages	0.242***	0.161***	0.141**	0.131**	0.199***	0.119**	0.056	0.051
	(0.058)	(0.060)	(0.057)	(0.059)	(0.057)	(0.059)	(0.056)	(0.058)
Effect of immigrants on own household's employment	-0.159**	-0.125**	0.106*	0.111*	-0.183***	-0.148**	0.069	0.084
	(0.062)	(0.064)	(0.060)	(0.062)	(0.061)	(0.063)	(0.059)	(0.061)
Effect of immigrants on most Americans' employment	0.370***	0.295***	0.258***	0.193***	0.295***	0.243***	0.210***	0.148***
	(0.054)	(0.055)	(0.055)	(0.057)	(0.053)	(0.055)	(0.053)	(0.055)
Effect of immigrants on culture					0.253***	0.200***	0.284***	0.263***
					(0.036)	(0.037)	(0.037)	(0.038)
Observations	771	771	723	723	771	771	723	723

Notes: This table uses data from respondents in the control group. The table shows OLS regression results where the dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. The independent variables are beliefs about the impact of low-skilled and high-skilled immigrants on the wages and employment of (i) people's own household and (ii) most Americans. On top of that people's beliefs about the effect of immigrants on American culture are included on columns (5) to (8). All questions were answered on 5-point Likert scales where higher values indicate more optimistic views regarding the effect of immigration. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.12: Assessing relative importance of cultural vs. economic concerns

	Low-skilled		High-skilled	
	Not familiar	Familiar	Not familiar	Familiar
Overall economic impact on most Americans	0.580*** (0.029)	0.488*** (0.031)	0.477*** (0.031)	0.427*** (0.034)
Overall cultural impact on American culture and society	0.161*** (0.029)	0.131*** (0.031)	0.216*** (0.031)	0.196*** (0.034)
Observations	762	762	714	716

Notes: This table uses data from respondents in the control group. We use the same controls as described in Table 1.1. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses.

Table A.13: Response to demand treatments

	Pro-immigration:		
	Main study	Main study - with follow-up sample	follow-up study
Demand treatment	0.220*** (0.078)	0.202* (0.103)	0.044 (0.108)
Observations	555	340	340

Notes: This table uses data from MTurk respondents. Pro-immigrant is people's support for immigration (z-scored) as measured by their support for low-skilled and high-skilled immigration. 'Demand treatment' takes value 1 for respondents who received the message that 'we are a group of researchers in favour of increasing immigration', and zero otherwise. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses.

1.D Summary statistics, balance and attrition

Table A.14: Characteristics of our sample compared to the US Census

	Mean: Online sample	Mean: Online sample: follow-up	Mean: ACS
Gender	0.48	0.50	0.51
What is your age?	46.6	47.5	47.1
Northeast	0.22	0.23	0.179
Midwest	0.18	0.19	0.211
West	0.24	0.23	0.24
Total household income	72907	73516	82433

Notes: This table summarizes the characteristics of our sample in the main survey as well as the follow-up survey along targeted dimensions as well as the characteristics of the 2015 American Community Survey.

Table A.15: Summary statistics

	Mean	SD	Median	Min.	Max.	Obs.
Gender	0.48	0.50	0.00	0.00	1.00	3130
What is your age?	46.64	15.59	49.50	21.00	69.50	3130
Non-Hispanic White	0.46	0.50	0.00	0.00	1.00	3130
African American/Black	0.06	0.23	0.00	0.00	1.00	3130
Republican	0.25	0.43	0.00	0.00	1.00	3130
Democrat	0.37	0.48	0.00	0.00	1.00	3130
Independent	0.36	0.48	0.00	0.00	1.00	3130
Northeast	0.22	0.41	0.00	0.00	1.00	3130
Midwest	0.18	0.39	0.00	0.00	1.00	3130
West	0.24	0.43	0.00	0.00	1.00	3130
Household size	2.39	1.48	2.00	0.00	10.00	3130
Total household income	72899.36	50744.30	62500.00	7500.00	225000.00	3130
Education	4.94	2.06	5.00	0.00	9.00	3130
High-skilled	0.75	0.43	1.00	0.00	1.00	3130
Full-time employee	0.46	0.50	0.00	0.00	1.00	3130
Part-time employee	0.08	0.27	0.00	0.00	1.00	3130
Self-employed or small business owner	0.06	0.23	0.00	0.00	1.00	3130
Unemployed and looking for work	0.05	0.21	0.00	0.00	1.00	3130
Student	0.03	0.16	0.00	0.00	1.00	3130
Not in labor force	0.26	0.44	0.00	0.00	1.00	3130
College	0.78	0.41	1.00	0.00	1.00	3130
Beliefs about effect of immigration on wages of low-skilled workers	2.21	1.00	2.00	1.00	5.00	3130
Beliefs about effect of immigration on wages of high-skilled workers	2.97	0.85	3.00	1.00	5.00	3130
Beliefs about effect of immigration on unemployment of low-skilled workers	2.10	1.02	2.00	1.00	5.00	3130
Beliefs about effect of immigration on unemployment of high-skilled workers	2.90	0.80	3.00	1.00	5.00	3130
Confidence in beliefs about wages of low-skilled workers	3.49	0.98	3.00	1.00	5.00	3130
Confidence in beliefs about unemployment of low-skilled workers	3.55	0.96	4.00	1.00	5.00	3130
Confidence in beliefsabout wages of high-skilled workers	3.52	0.97	3.00	1.00	5.00	3130
Confidence in beliefs about unemployment of high-skilled workers	3.50	0.96	3.00	1.00	5.00	3130

Notes: This table displays the summary statistics for our sample.

Table A.16: Balance across the treatment and control group

	Treatment	Control	P-value(Treatment - Control)	Observations
Gender	0.48	0.48	0.945	3130
What is your age?	45.84	46.54	0.223	3130
Non-Hispanic White	0.47	0.45	0.171	3130
African American/Black	0.05	0.06	0.093	3130
Republican	0.26	0.25	0.543	3130
Democrat	0.37	0.38	0.407	3130
Independent	0.37	0.36	0.842	3130
Northeast	0.23	0.22	0.356	3130
Midwest	0.18	0.18	0.958	3130
West	0.23	0.25	0.258	3130
Household size	2.53	2.59	0.235	3130
Log household income	10.91	10.91	0.958	3130
High-skilled	0.76	0.75	0.675	3130
Full-time employee	0.50	0.48	0.409	3130
Part-time employee	0.09	0.08	0.411	3130
Self-employed or small business owner	0.06	0.07	0.279	3130
Unemployed and looking for work	0.05	0.05	0.504	3130
Student	0.03	0.03	0.414	3130
Not in labor force	0.27	0.30	0.107	3130
College	0.85	0.83	0.084	3130
Beliefs about effect of immigration on wages of low-skilled workers	2.20	2.23	0.434	3130
Beliefs about effect of immigration on wages of high-skilled workers	2.96	2.97	0.624	3130
Beliefs about effect of immigration on unemployment of low-skilled workers	2.12	2.09	0.371	3130
Beliefs about effect of immigration on unemployment of high-skilled workers	2.90	2.91	0.824	3130
Confidence in beliefs about wages of low-skilled workers	3.43	3.52	0.016	3130
Confidence in beliefs about unemployment of low-skilled workers	3.51	3.57	0.078	3130
Confidence in beliefsabout wages of high-skilled workers	3.49	3.51	0.597	3130
Confidence in beliefs about unemployment of high-skilled workers	3.44	3.51	0.052	3130

Notes: This table examines covariate balance based on the sample of respondents from the main study. The p-value of an F-test testing for the joint significance of all covariates in predicting treatment status is 0.2302.

Table A.17: Balance across the treatment and control group in the follow-up

	Treatment	Control	P-value(Treatment - Control)	Observations
Gender	0.49	0.50	0.708	3130
What is your age?	46.92	47.97	0.126	3130
Non-Hispanic White	0.49	0.49	0.908	3130
African American/Black	0.05	0.07	0.019	3130
Republican	0.26	0.26	0.953	3130
Democrat	0.37	0.38	0.756	3130
Independent	0.36	0.35	0.752	3130
Northeast	0.23	0.23	0.979	3130
Midwest	0.18	0.20	0.133	3130
West	0.24	0.23	0.620	3130
Household size	2.46	2.51	0.450	3130
Log household income	10.93	10.92	0.782	3130
High-skilled	0.75	0.74	0.641	3130
Full-time employee	0.49	0.45	0.093	3130
Part-time employee	0.09	0.08	0.344	3130
Self-employed or small business owner	0.05	0.06	0.163	3130
Unemployed and looking for work	0.05	0.05	0.956	3130
Student	0.02	0.02	0.527	3130
Not in labor force	0.29	0.33	0.044	3130
College	0.84	0.83	0.759	3130
Beliefs about effect of immigration on wages of low-skilled workers	2.22	2.25	0.564	3130
Beliefs about effect of immigration on wages of high-skilled workers	2.95	2.95	0.888	3130
Beliefs about effect of immigration on unemployment of low-skilled workers	2.13	2.10	0.541	3130
Beliefs about effect of immigration on unemployment of high-skilled workers	2.90	2.91	0.786	3130
Confidence in beliefs about wages of low-skilled workers	3.42	3.51	0.040	3130
Confidence in beliefs about unemployment of low-skilled workers	3.50	3.55	0.213	3130
Confidence in beliefsabout wages of high-skilled workers	3.48	3.51	0.563	3130
Confidence in beliefs about unemployment of high-skilled workers	3.45	3.52	0.080	3130

Notes: This table examines covariate balance for the follow-up sample. The p-value of an F-test testing for the joint significance of all covariates in predicting treatment status is 0.4428.

Table A.18: Attrition in the follow-up study

	Completed Follow-up Survey	Completed Follow-up Survey
Treatment	-0.004	-0.002
	(0.016)	(0.017)
Republican	0.005	
	(0.021)	
Independent	-0.027	
	(0.019)	
Log(Income)	-0.022**	
	(0.011)	
High-skilled	-0.116***	
	(0.020)	
Employed Full-Time	-0.005	
	(0.023)	
Employed Part-Time	0.005	
	(0.035)	
Unemployed	0.049	
	(0.043)	
Self-Employed	-0.046	
	(0.039)	
Student	-0.083	
	(0.058)	
High Education	-0.046	
	(0.032)	
Male	0.041**	
	(0.017)	
Age	0.002***	
	(0.001)	
Observations	3130	3130

Notes: The outcome variables takes value 1 for respondents who completed the follow-up study. 'Treatment" is an indicator equal to 1 if respondents received the research evidence. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses.

1.D.1 Heterogeneous effects

Table A.19: Heterogeneous Treatment Effects: Self-reports

	Low-skilled			High-skilled		
	Not familiar	Familiar	Index	Not familiar	Familiar	Index
Panel A: Prior Belief						
Treatment × (a)	-0.057	-0.055	-0.056	0.029	0.022	0.026
Prior	(0.043)	(0.046)	(0.040)	(0.046)	(0.049)	(0.044)
Treatment (b)	0.169***	0.117***	0.143***	0.105***	0.045	0.075**
	(0.033)	(0.034)	(0.030)	(0.033)	(0.034)	(0.030)
Pr(a+b)=0	0.037	0.265	0.079	0.014	0.234	0.046
Observations	3130	3130	3130	3130	3130	3130
Panel B: Republican						
Treatment × (a)	-0.052	0.011	-0.021	0.010	-0.016	-0.003
Republican	(0.075)	(0.082)	(0.071)	(0.078)	(0.083)	(0.072)
Treatment (b)	0.182***	0.115***	0.149***	0.102***	0.049	0.075**
	(0.038)	(0.038)	(0.035)	(0.038)	(0.038)	(0.034)
Pr(a+b)=0	0.047	0.087	0.040	0.102	0.654	0.256
Observations	3130	3130	3130	3130	3130	3130
Panel C: High-skill						
Treatment × (a)	-0.039	-0.051	-0.045	0.089	-0.010	0.039
High-skill	(0.075)	(0.078)	(0.069)	(0.076)	(0.080)	(0.070)
Treatment (b)	0.198***	0.156**	0.177***	0.038	0.052	0.045
	(0.065)	(0.067)	(0.059)	(0.066)	(0.069)	(0.060)
Pr(a+b)=0	0.000	0.009	0.000	0.001	0.280	0.017
Observations	3130	3130	3130	3130	3130	3130

Notes: The table shows OLS regression results where the dependent variables are attitudes to the four different types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions, (iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. The indices are created by taking the mean of the responses to immigrants with different familiarity with American values and traditions for each skill level. "Treatment" is an indicator equal to 1 if respondents received the research evidence. "Prior" is people's pre-treatment belief about the wage and employment effects of the Mariel boatlift on low-skilled (high-skilled) workers. "Republican" takes value 1 if our respondent self-identifies as a Republican and zero otherwise. "High-skill" takes values 1 if our respondent self-identifies as high-skilled and zero otherwise. We use the same controls as described in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.20: Heterogeneous Treatment Effects: Petition

	Intention: H2B Visas		
	Increase	Decrease	Net support
Panel A: Prior Belief			
Treatment × (a)	-0.006	0.015	-0.035
Prior	(0.019)	(0.019)	(0.042)
Treatment (b)	0.046***	-0.058***	0.134***
	(0.016)	(0.016)	(0.034)
Pr(a+b)=0	0.124	0.062	0.065
Observations	3130	3130	3130
Panel B: Republican			
Treatment × (a)	-0.025	-0.007	-0.021
Republican	(0.033)	(0.038)	(0.076)
Treatment (b)	0.052***	-0.056***	0.139***
	(0.019)	(0.017)	(0.040)
Pr(a+b)=0	0.314	0.066	0.069
Observations	3130	3130	3130
Panel C: High-skill			
Treatment × (a)	-0.004	-0.016	0.034
High-skill	(0.035)	(0.036)	(0.079)
Treatment (b)	0.049	-0.046	0.108
	(0.030)	(0.032)	(0.068)
Pr(a+b)=0	0.016	0.001	0.000
Observations	3130	3130	3130

Notes: The three first columns show regression results where the dependent variable is intention to sign the petitions. "Increase" ("Decrease") is an indicator equal to 1 if a respondent wanted to sign the petition suggesting to increase (decrease) the annual cap on the H-2B visa program. "Net support" is a z-scored transformation of a variable taking value 1 (-1) if a respondent wanted to sign the petition to increase (decrease) the annual cap on the H-2B visa program and 0 otherwise. "Treatment" is an indicator equal to 1 if respondents received the research evidence. "Prior" is people's pre-treatment belief about the wage and employment effects of the Mariel boatlift on low-skilled (high-skilled) workers. "Republican" takes value 1 if our respondent self-identifies as a Republican and zero otherwise. "High-skill" takes values 1 if our respondent self-identifies as high-skilled and zero otherwise. We use the same controls as described in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.21: Heterogeneous Treatment Effects: Follow-up study

	Labor Market Concerns	Low-skilled Immigrants	High-skilled Immigrants
Panel A: Prior Belief			
Treatment × (a)	-0.101**	-0.062	-0.037
Prior	(0.051)	(0.050)	(0.043)
Treatment (b)	0.115***	0.095**	0.058
	(0.040)	(0.040)	(0.041)
Pr(a+b)=0	0.820	0.598	0.716
Observations	2075	2075	2075
Panel B: Republican			
Treatment × (a)	0.054	0.070	0.018
Republican	(0.113)	(0.095)	(0.100)
Treatment (b)	0.137**	0.077*	0.054
	(0.062)	(0.045)	(0.046)
Pr(a+b)=0	0.044	0.082	0.419
Observations	2075	2075	2075
Panel C: High-skill			
Treatment × (a)	-0.152	-0.007	0.003
High-skill	(0.119)	(0.090)	(0.094)
Treatment (b)	0.265***	0.100	0.057
	(0.102)	(0.076)	(0.081)
Pr(a+b)=0	0.064	0.048	0.222
Observations	2075	2075	2075

Notes: The dependent variables are attitudes towards the two types of immigrant we asked about in the follow-up: (i) "workers with little to no education" and (ii) "highly educated workers". Responses were given on a five point scale indicating that immigration of these workers should be "strongly decreased" to "strongly increased". Respondents were also asked to what extent they agreed with the following statement: "Immigrants hurt American workers," which was answered on a five point scale from 1: "strongly disagree" to 5: "strongly agree" (we have reversed the scale). All responses are z-scored using the mean and standard deviation in the control group. "Treatment" is an indicator equal to 1 if respondents received the research evidence. "Prior" is people's pre-treatment belief about the wage and employment effects of the Mariel boatlift on low-skilled (high-skilled) workers. "Republican" takes value 1 if our respondent self-identifies as a Republican and zero otherwise. "High-skill" takes values 1 if our respondent self-identifies as high-skilled and zero otherwise. We use the same controls as described in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.22: Heterogeneous Treatment Effects by pre-treatment beliefs: labour Market Concerns

	Wages		Employment	
	Own household	Most Americans	Own household	Most Americans
Panel A: Effect of low-skilled				
Treatment × Prior	-0.121**	-0.150***	-0.075	-0.165***
	(0.062)	(0.057)	(0.057)	(0.055)
Treatment	0.304**	0.510***	0.175	0.494***
	(0.146)	(0.134)	(0.132)	(0.125)
Observations	1489	1489	1484	1484
Panel B: Effect of high-skilled				
Treatment × Prior	-0.073	-0.038	-0.070	0.045
	(0.074)	(0.069)	(0.079)	(0.076)
Treatment	0.277	0.293	0.356	0.121
	(0.229)	(0.210)	(0.241)	(0.232)
Observations	1487	1487	1481	1481

Notes: The dependent variables are beliefs about the economic impact of low-skilled and high-skilled immigrants. Respondents were asked how they thought admitting more low-skilled/high-skilled immigrants would affect (i) "wages," (ii) "job opportunities and job security,". We randomised whether respondents answered these questions with respect to low-skilled or high-skilled immigrants. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views regarding the effect of immigration. "Treatment" is an indicator equal to 1 if respondents received the research evidence. "Prior" is people's pre-treatment belief about the wage and employment effects of the Mariel boatlift on low-skilled (high-skilled) workers. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.23: Heterogeneous Treatment Effects by Self-perceived Skill-level: labour Market Concerns

	Wages		Employment		Fiscal burden		Culture
	Own household	Most Americans	Own household	Most Americans	Own household	Most Americans	
Panel A: Effect of low-skilled							
Treatment ×	0.082	-0.095	-0.082	-0.151	-0.090	-0.022	-0.036
High-skilled	(0.125)	(0.115)	(0.125)	(0.117)	(0.119)	(0.117)	(0.118)
Treatment	-0.025	0.251**	0.082	0.262**	0.095	0.046	0.061
	(0.114)	(0.103)	(0.113)	(0.106)	(0.106)	(0.105)	(0.103)
Observations	1489	1489	1484	1484	1479	1479	1469
Panel B: Effect of high-skilled							
Treatment ×	0.103	0.024	0.063	0.078	0.035	-0.026	0.209*
High-skilled	(0.102)	(0.106)	(0.107)	(0.107)	(0.108)	(0.108)	(0.108)
Treatment	-0.012	0.167*	0.108	0.200**	0.028	0.094	-0.111
	(0.087)	(0.091)	(0.091)	(0.090)	(0.093)	(0.094)	(0.092)
Observations	1487	1487	1481	1481	1475	1475	1461

Notes: The dependent variables are beliefs about the economic impact of low-skilled and high-skilled immigrants. Respondents were asked how they thought admitting more low-skilled/high-skilled immigrants would affect (i) "wages," (ii) "job opportunities and job security," and (iii) "taxes" for their own household as well as for most Americans. We randomised whether respondents answered these questions with respect to low-skilled or high-skilled immigrants. All questions were answered on 5-point Likert scales where higher values indicate more optimistic views regarding the effect of immigration. "Treatment" is an indicator equal to 1 if respondents received the research evidence. "High-skill" takes values 1 if our respondent self-identifies as high-skilled and zero otherwise. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

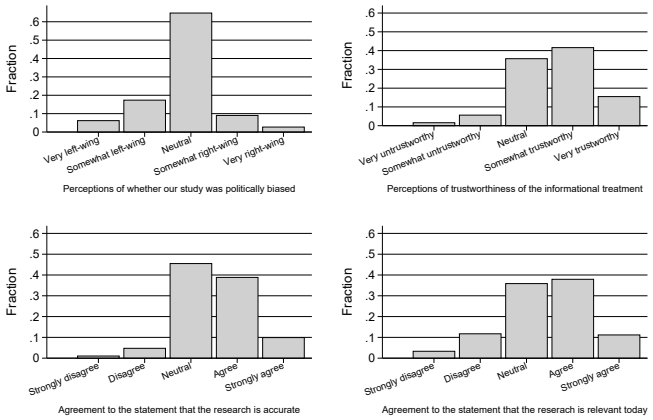
Table A.24: Mechanisms: Beliefs about the overall economic impact of immigrants (post-treatment)

	Low-skilled		High-skilled		Follow-up
	Not familiar	Familiar	Not familiar	Familiar	All Immigrants
Panel A: Main					
Treatment	0.218***	0.172***	0.151***	0.134***	0.122***
	(0.048)	(0.049)	(0.048)	(0.048)	(0.040)
Panel B: Prior					
Treatment ×	-0.022	-0.044	0.104	0.112*	-0.115*
Prior	(0.064)	(0.067)	(0.069)	(0.067)	(0.065)
Treatment	0.218***	0.172***	0.152***	0.135***	0.117***
	(0.048)	(0.049)	(0.048)	(0.048)	(0.039)
Panel C: Republican					
Treatment ×	-0.013	0.030	0.016	0.045	0.041
Republican	(0.112)	(0.113)	(0.110)	(0.115)	(0.086)
Treatment	0.221***	0.164***	0.150***	0.119**	0.104**
	(0.054)	(0.056)	(0.056)	(0.054)	(0.047)
Observations	1472	1472	1472	1468	2075

Notes: The table shows OLS regression results where the dependent variables are beliefs about the overall economic impact on most Americans of the four different types of immigrants. Beliefs were elicited using the following question: "When you think about all of the potential positive and negative economic effects of increasing the number of high-skilled/low-skilled immigrants that are not familiar/highly familiar with American values and traditions coming to the United States, do you think the overall effect would be positive or negative for the finances of most Americans?" The question was answered on a five point scale from 1: "The overall economic effect would be very negative for most Americans" to 5: "The overall economic effect would be very positive for most Americans". We randomised whether respondents were asked these questions about low-skilled immigrants or high-skilled immigrants. The outcome variables are z-scored using the mean and standard deviation in the control group. We use the same controls as described in Table 1.1. * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors in parentheses.

1.E Beliefs about the survey and the research evidence

Figure A.6: Beliefs about the survey and the research evidence



Notes: The figure displays results on a question measuring people's perception of the trustworthiness in the research evidence. The figure shows the fraction of respondents who think that the survey was (1) very untrustworthy, (2) untrustworthy, (3) neither untrustworthy nor trustworthy, (4) trustworthy, and (5) very trustworthy.

Table A.25: Beliefs about the research study and attitudes towards immigrants

	Index: Attitudes towards	
	low-skilled immigrants	high-skilled immigrants
Info: trustworthy	0.179***	0.167***
	(0.036)	(0.037)
Study: internally valid	0.073*	0.059
	(0.040)	(0.041)
Study: externally valid	0.172***	0.145***
	(0.028)	(0.028)
Observations	1451	1451

Notes: This table uses data from respondents in the treatment group. The table shows OLS regression results where the dependent variables are indices of attitudes towards lows-skilled and high-skilled immigrants. The index on low-skilled immigrants is based on the following two two types of immigrants: (i) low-skilled immigrants not familiar with American values and traditions, (ii) low-skilled immigrants highly familiar with American values and traditions. The index on high-skilled immigrants is based on the following two two types of immigrants:(iii) high-skilled immigrants not familiar with American values and traditions, and (iv) high-skilled immigrants highly familiar with American values and traditions. The answers were given on a five point scale from 1: "Allow a lot less of these immigrants" to 5: "Allow a lot more of these immigrants". The question order was randomised (statistical tests show no order effects). The outcomes are z-scored using the mean and standard deviation in the control group. The main independent variables in these regressions are people's assessment of whether the research evidence we provided them with is (i) trustworthy, (ii) reflective of the actual labour market impacts of the Mariel boatlift, and (iii) relevant for assessing the costs and benefits of allowing more or less immigrants into the US today. The questions were asked on 5-point Likert scales, where higher values indicate more positive views on the research evidence. We use the same controls as described in Table 1.1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

Table A.26: Correlates of beliefs about the research study

	The information is	The research study is	
	trustworthy	internally valid	externally valid
Republican	-0.391***	-0.360***	-0.480***
	(0.059)	(0.053)	(0.064)
Independent	-0.302***	-0.262***	-0.357***
	(0.052)	(0.046)	(0.055)
Age	-0.004**	-0.005***	-0.007***
	(0.002)	(0.002)	(0.002)
Log(Income)	-0.005	-0.014	0.014
	(0.030)	(0.028)	(0.033)
High-skilled	0.088	0.071	0.058
	(0.059)	(0.053)	(0.062)
Employed Full-Time	-0.122*	-0.033	-0.056
	(0.063)	(0.056)	(0.068)
Employed Part-Time	-0.132	-0.119	-0.056
	(0.089)	(0.080)	(0.092)
Unemployed	-0.243**	-0.122	-0.030
	(0.115)	(0.116)	(0.123)
Self-Employed	-0.202*	-0.182*	-0.253**
	(0.122)	(0.104)	(0.123)
Student	-0.053	-0.027	0.074
	(0.155)	(0.154)	(0.155)
High Education	-0.041	-0.026	-0.007
	(0.086)	(0.079)	(0.089)
Male	0.014	0.041	0.022
	(0.045)	(0.041)	(0.050)
Observations	1458	1456	1451

Notes: This table uses data from respondents in the treatment group. The main dependent variables in this regressions are people's assessment of whether the research evidence we provided them with is (i) trustworthy, (ii) reflective of the actual labour market impacts of the Mariel boatlift, and (iii) relevant for assessing the costs and benefits of allowing more or less immigrants into the US today. The questions were asked on 5-point likert scales, where higher values indicate more positive views on the research evidence. * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors in parentheses.

1.F Miscellaneous

Figure A.7: Petition

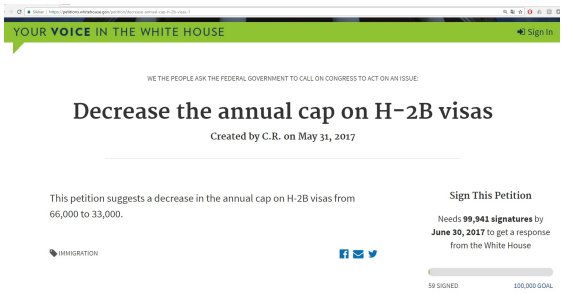


Figure A.8: Invitation form in the email sent out for the obfuscated follow-up study

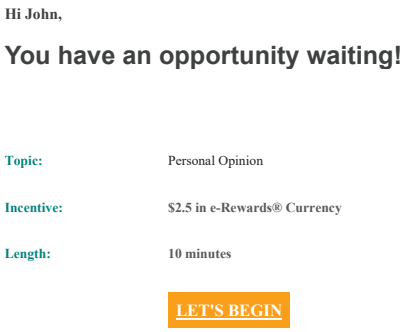


Figure A.9: Consent form in the main study

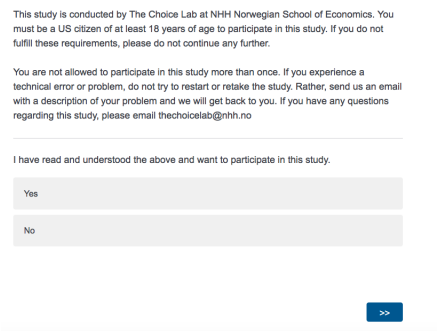


Figure A.10: Consent form in the follow-up study

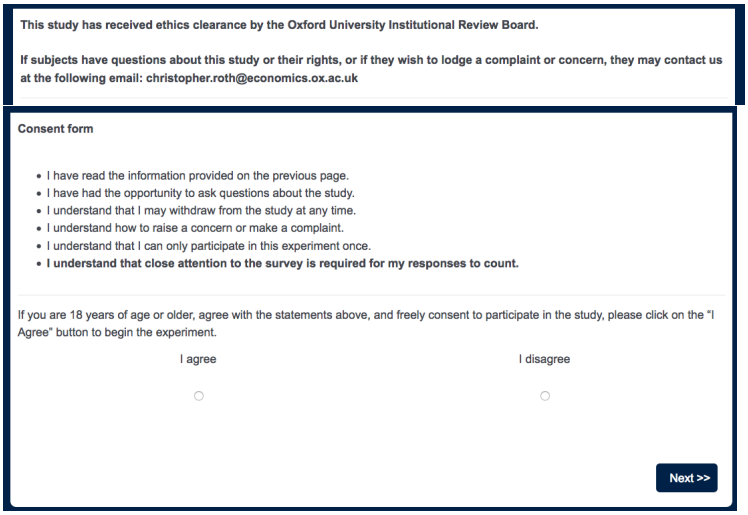


Figure A.11: Screenshot of the informational treatment

NHH

The researchers who analyzed the short- and long-term effects of the mass immigration of Cubans to Miami concluded that, for both high-skilled and low-skilled workers, the mass immigration had **virtually no effect on wages** and **virtually no effect on unemployment**.

According to the researchers, the mass immigration had virtually no effect on wages and unemployment because the new Cuban immigrants increased the overall demand for goods and services, which created more jobs.

Chapter 2

Measuring and Bounding Experimenter Demand

Jonathan de Quidt, Johannes Haushofer, and Christopher Roth[1]

Forthcoming, American Economic Review

[1]de Quidt: Institute for International Economic Studies and CESifo, address: Institute for International Economic Studies, Stockholm University, 106 91 Stockholm, Sweden, email: jonathan.dequidt@iies.su.se. Haushofer: Department of Psychology, Princeton University, NBER, and Busara Center for behavioural Economics, address: 427 Peretsman-Scully Hall, Princeton University, Princeton, NJ 08540, United States, email: haushofer@princeton.edu. Roth: Department of Economics, University of Oxford, and CSAE, address: Keble College, Parks Road, Oxford, OX1 3PG, United Kingdom, email: christopher.roth@economics.ox.ac.uk. We thank the editors, Stefano DellaVigna and Esther Duflo, and three anonymous referees for useful comments. We are grateful to Johannes Abeler, Dan Benjamin, Stefano Caria, Rachel Cassidy, Tom Cunningham, Elwyn Davies, Armin Falk, Thomas Graeber, Don Green, Alexis Grigorieff, Johannes Hermle, George Loewenstein, Simon Quinn, Matthew Rabin, Gautam Rao, Bertil Tungodden and Liad Weiss for comments. We thank Stefano DellaVigna, Lukas Kiessling, and Devin Pope for sharing code. Moreover, we thank various seminar participants. We thank Justin Abraham for excellent research assistance. de Quidt acknowledges financial support from Handelsbanken's Research Foundations, grant no: B2014-0460:1. The experiments in this paper were funded by Princeton University. IRB approval was obtained at Princeton University and the University of Oxford. The experiments were preregistered in trial 1248 on the American Economic Association RCT Registry, available at https://www.socialscienceregistry.org/trials/1248.

Abstract We propose a technique for assessing robustness to demand effects of findings from experiments and surveys. The core idea is that by deliberately inducing demand in a structured way we can bound its influence. We present a model in which participants respond to their beliefs about the researcher's objectives. Bounds are obtained by manipulating those beliefs with "demand treatments". We apply the method to eleven classic tasks, and estimate bounds averaging 0.13 standard deviations, suggesting that typical demand effects are probably modest. We also show how to compute demand-robust treatment effects and how to structurally estimate the model.

2.1 Introduction

A basic concern in experimental work with human participants is that, knowing that they are being experimented on, the participants may change their behavior. Specifically, participants may try to infer the experimenter's objective from their treatment, and then act accordingly (Orne, 1962; Rosenthal, 1966; Zizzo, 2010). For instance, participants who believe the researcher wants to show that people free-ride in public good games might play more selfishly than they otherwise would. Thus, instead of measuring the participant's "natural" choice, the data are biased by an unobservable *experimenter demand effect*. Demand effects pose a threat to external validity, because participants would make different choices if the experimenter were absent. They can affect estimates of average behaviour and treatment effects, and have been raised as a concern in the context of lab experiments (Levitt and List, 2007; List, 2006; List et al., 2004), field experiments (Al-Ubaydli et al., 2017; Allcott and Taubinsky, 2015; Dupas and Miguel, 2017), and survey responses (Bertrand and Mullainathan, 2001; Clark and Schober, 1992).[2]

[2]Zizzo (2010) discusses how demand effects can arise from different sources, such as perceived social pressure from the experimenter, or inferences about appropriate behav-

The core idea of our paper is that one can construct plausible bounds on demand-free behaviour and treatment effects by deliberately *inducing* experimenter demand and measuring its influence. For example, in a dictator game, we explicitly tell some participants that we expect they will give more than they normally would, while others are told we expect they will give less. Under the assumption that any underlying demand effect is less extreme than our manipulations (in a sense that we will formalize), choices under these instructions give upper and lower bounds on demand-free behavior, and by combining bounds from different experimental treatments we can estimate bounds on treatment effects.

We begin with a simple Bayesian model of decision-making that motivates our approach. In our model, an experiment defines a mapping from actions to utility. The experimenter is only interested in measuring the "natural" action (or changes in that action) that maximizes the participant's utility as derived from the experimental payoffs. However, the participant is also motivated to take actions that conform to the experimenter's research objectives. He infers those objectives from the design features, and distorts his action, biasing the results. Our demand treatments manipulate those beliefs to identify an interval containing the natural action. We remain agnostic about *why* the participant wishes to please the experimenter; motives could include altruism, a desire to conform, a misguided attempt to contribute to science, or an expectation of reciprocity from the experimenter.

ior. In psychology, experimenter demand effects are considered a specific case of "demand characteristics" (Orne, 1962), which also include the simple effect of being observed ("Hawthorne" effects), or the effect of features of the environment on task construal. Researchers might also worry about "social desirability bias" (respondents taking actions they perceive to be moral or desirable, which may or may not relate to the researcher's objectives), or responses motivated by respondents' own preferences over the findings (e.g. respondents might misreport income in a survey to increase their eligibility for a program). In this paper we focus on inferences about the experimenter's objective, but the framework can easily be adapted to fit other inferences.

We provide an extensive set of applications of the method. We conduct seven online experiments with approximately 19,000 participants in total, in which we construct bounds on demand-free behaviour for 11 canonical tasks.[3] We employ two different types of demand treatments. "Weak" demand treatments signal an experimental hypothesis to our respondents: we tell them "We expect that participants who are shown these instructions will [work, invest, ...] more/less than they normally would". We believe that these treatments are likely to be more informative than implicit signals about demand in typical studies, so in our view these bounds will be sufficient for most applications. Our "strong" demand treatments go further, telling participants "You will do us a favour if you [work, invest, ...] more/less than you normally would".[4] These give rise to much more conservative bounds, which may be useful for applications where concerns about demand are paramount. They also play an important role in our more structural applications, described below, and their strength makes them suited for studying demand effects in their own right.

We establish several novel facts about demand effects. Our first finding is that responses to the weak treatments are modest, averaging around 0.13 standard deviations, varying from close to zero for unincentivised real effort to 0.29 standard deviations for trust game second movers. In most tasks, our estimates are not significantly different from zero. Overall, we interpret these results as suggesting that demand effects in typical experiments are

[3]Specifically, we study simple time, risk and ambiguity preference elicitation tasks, a real effort task with and without performance incentives, a lying game, dictator game, ultimatum game (first and second mover), and trust game (first and second mover).

[4]We based this phrasing on Binmore et al.'s (1985) experiment on the ultimatum game, in which the instructions included the line "You will be doing us a favour if you simply set out to maximize your winnings". These instructions were subsequently criticised precisely because they potentially induce experimenter demand (see e.g. Zizzo, 2010). In recent work, Ellingsen et al. (2018) use similar language, deliberately using demand to try to shut down social preference motivations in games with communication.

likely to be small. Responses to our strong demand treatments are much larger, with bounds averaging 0.6 standard deviations and ranging from 0.23 to 1.06 standard deviations. While these bounds are likely more conservative than required in most applications, they illustrate that participants can respond substantially to strong signals about the researcher's objective, thus researchers are right to pay close attention to potential demand effects in their studies.

The heterogeneity across tasks in responsiveness to our treatments reveals differing levels of *uncertainty* about the importance of experimenter demand in different tasks. For example, there is more uncertainty (i.e., wider bounds) about demand effects for trust game second movers than in the effort task. We provide an additional assumption, "monotone sensitivity," under which this heterogeneity can be interpreted as revealing variation in the *magnitude* of demand effects in different tasks, i.e. that demand effects are larger for trust game second movers.

Next, we apply the method to bounding treatment effect estimates, deriving bounds on the real effort response to performance pay. The bounds we obtain using our weak demand treatments are quite tight, corresponding to around 11 percent of the estimated treatment effect (or 0.07 standard deviations). The strong demand treatments generate wider bounds, but even these more conservative bounds exclude zero, supporting the qualitative finding that incentives increase effort. We apply standard methods to construct "demand-robust" confidence intervals on the bounds and on the underlying actions or treatment effects contained by those bounds. These intervals combine the standard parameter uncertainty due to sampling error with the additional uncertainty due to potential demand effects.

Third, we turn to point estimation of treatment effects. We ask whether applying same-signed demand treatments to both the control and treatment group (for example, demanding high effort from both groups) can reduce or eliminate bias due to experimenter demand. Intuitively, the goal is to "control for" demand by harmonizing beliefs across treatments. We show that this approach is valid under additional assumptions, and apply it to the effort experiment, obtaining a set of alternative estimates, all lying within 10 percent of the conventional treatment effect estimate.

Fourth, following the basic approach of DellaVigna and Pope (2018), we illustrate how sufficiently informative demand treatments can be used in conjunction with a structural model to obtain unconfounded estimates of structural parameters of interest and measure participants' value of conforming to the experimenter's wishes. We estimate that the value of pleasing the experimenter in our effort task is equivalent to increasing the monetary incentives by 20 percent.

Fifth, we explore some of the properties of demand effects. Our approach relies on a Monotonicity assumption, essentially assuming that participants want to comply with rather than defy the researchers' wishes. We find strong support for this assumption in average behavior, and at the individual level, using a within-participants design. We show using simple belief data that participants' beliefs about the experimental objective respond as expected to our demand treatment. We also compare our bounds to estimates of the effect of double anonymity in dictator games, one manipulation that has been interpreted as reducing demand. Finally, we examine four moderators of sensitivity to experimenter demand: incentivised versus hypothetical choice; gender; attention; and participant pool.

Finally, we provide an extended summary of recommendations for practitioners, covering how to apply the methods developed and practical lessons learned from our own applications.

We contribute to the small literature discussing experimenter demand effects (Fleming and Zizzo, 2014; Shmaya and Yariv, 2016; Zizzo, 2010), demand characteristics (Orne, 1962), and obedience to the experimenter (Milgram, 1963). We are aware of few attempts to directly assess the empirical importance of experimenter demand, and a key contribution of our paper is to provide a general framework for studying demand effects and evidence from a wide range of standard tasks. In recent work, concurrent with our own, Mummolo and Peterson (2017) conduct two vignette studies on support for free speech and partisan news consumption, and a hypothetical audit study concerning racial bias in hiring, using treatments similar to our weak demand treatments.[5] While they do not construct bounds, they find modest responses to these treatments, in line with our findings.[6]

Relatedly, our paper contributes to the literature on social pressure (DellaVigna et al., 2012, 2017) and moral suasion (Dal Bó and Dal Bó, 2014).

We also relate to the literature which examines the effects of anonymity on

[5]For example, some participants in the audit study are told "We expect that job candidates with names indicating they are white will be more likely to receive an interview because of the historical advantages this group has had on the job market," while others are told "We expect that job candidates with names indicating they are African American will be more likely to receive an interview because corporations are increasingly looking to diversify their workforces".

[6]Other related papers include Cilliers et al. (2015), who show that a white foreigner's presence in the lab in experiments in Sierra Leone distorted giving in dictator games; Lambdin and Shaffer (2009), who find that participants' ability to guess hypotheses varied (but was mostly low) across three different experimental tasks; Bischoff and Frank (2011), in which an actor (unsuccessfully) tried to induce demand effects by their delivery of instructions in a lab game; and Tsutsui and Zizzo (2014) who measure individual demand sensitivity by participants' propensity to select dominated lotteries from a list when told "it would be nice if some of you were to choose" them. List (2007) and Bardsley (2008) argue that behaviour in the dictator game is to a large degree an artifact of the experimental situation.

behaviour in the laboratory. Participants who believe their choices are being monitored might be more likely to try to please the experimenter. Hoffman et al. (1994) and List et al. (2004) find that varying anonymity can influence pro-social behavior, while Barmettler et al. (2012) find little effect. Intriguingly, Loewenstein (1999) suggests that participants' responses to the anonymity treatments in Hoffman et al. (1994) could themselves be driven by demand. Our findings also complement work that explores the principal-agent relationship between experimenter and participant (Chassang et al., 2012; Shmaya and Yariv, 2016).

Finally, our paper relates to the debate on how lab behaviour generalizes to the field (Camerer, 2012; Falk and Heckman, 2009; Harrison and List, 2004; Levitt and List, 2007; List, 2006). There are multiple reasons why behaviour might differ between lab and field, including demand effects. Our focus is on bounding the influence of demand while holding constant other design features. In some cases there may exist a "natural field experiment" counterpart to the design of interest, in which participants are unaware of the experiment, addressing demand alongside other external validity concerns. However, the set of studies that can be practically conducted as natural field experiments is limited. This literature often highlights a distinction between *qualitative* (directional) and *quantitative* effects. Either could be threatened by experimenter demand. Our approach can be used to put quantitative bounds on point estimates, but also to assess whether a qualitative finding could be explained by a demand effect, for instance by asking whether the bounds exclude zero or a sign reversal.

One indication of the level of concern about demand is the consideration given to it in study design. The experimental toolbox contains a number

of techniques that are partly or wholly motivated by the goal of reducing the influence of experimenter demand. For example, researchers often work hard to conceal potential signals about the study objective (such as efforts to avoid making gender salient, Bordalo et al., 2016); favour between-participant designs despite the larger samples required (Charness et al., 2012);[7] or conduct costly natural field experiments (Harrison and List, 2004).[8] These approaches plausibly make it more difficult for participants to infer the true experimental hypothesis – hopefully reducing the correlation between inference and treatment – or reduce participants' responsiveness to their inferences. But it is difficult to be sure that one has been successful, or that participants are not acting out some other conjecture that could be correlated in unpredictable ways with treatment. It is also difficult to know what is the set of studies that remains unpublished, or not even conducted, due to unresolved concerns about demand. Our bounding approach seeks to isolate the hidden demand effects by *amplifying* them with an explicit demand effect. It can be applied broadly without requiring major changes to experimental design, and we believe it will prove a useful addition to the toolbox.

The paper proceeds as follows. Section 2.2 presents a simple model of experimenter demand. Section 2.3 describes the experiments. Section 2.4 presents bounds on natural actions and treatment effects, demand-corrected point estimates, and structural estimates. Section 2.5 examines properties of demand effects and the assumptions underlying our approach. Section 2.6

[7] Charness et al. (2012) write (p2), "Within designs may lead to spurious effects, through respondents expecting to act in accord with some pattern, or attempting to provide answers to satisfy their perceptions of the experimenter's expectations... Demand effects are likely to be stronger in a within design".

[8] Other design features include abstract framing of choices, anonymised responses, homogenised delivery of instructions and incentivised choice. Review articles by Zizzo (2010) and de Quidt et al. (2018) provide a discussion, de Quidt et al. (2018) also measure their adoption in published experimental papers.

provides guidance for applying our approach in different settings. Section 2.7 concludes. A set of appendices contains theoretical details and additional results.

2.2 Theory

We now derive a simple model of experimenter demand and demand treatments. We begin with the three central assumptions at the heart of our approach, and provide a Bayesian model that generates them. Next we discuss demand treatment design. We conclude with a brief discussion of heterogeneity, and defiers, participants who do the opposite of the experimenter's wishes. Appendices 2.B.5 and 2.B.6 extend the model to allow participants to infer the *importance* of the experimenter's objective, and to model demand treatments that ask participants to ignore the experimenter's objective.

We model a decision-maker (he) who has preferences over outcomes induced by his action $a \in \mathbb{R}$ in an experiment. a could be continuous or discrete, but for simplicity we focus on the case of continuous actions with a natural ordering (more/less effort, investment, giving).

In the absence of demand effects, the optimal action is simply a function of the decision-making *environment*. We index environments by $\zeta \in Z$, where ζ captures aspects including participant characteristics (e.g. male/female, student/representative sample), setting (e.g. lab/field, online/in-person), experimental treatments, the content and framing of information provided to participants, and so on. A key component of ζ is information the participant has about other treatments (e.g. in a within-participant design), which might inform their beliefs about the experimental objective.

Given ζ, we define the "natural" action $a(\zeta)$ as that which would be taken absent any confounding motive for pleasing the experimenter.[9] The experimenter (she) is interested in measuring a specific action $a(\zeta)$ (e.g., the level of giving out of an endowment), or a treatment effect $a(\zeta_1) - a(\zeta_0)$ (e.g., the effect of incentives on effort provision). Unfortunately, her task is complicated by experimenter demand. After observing ζ, the decision-maker forms a conjecture about the experimenter's wishes or objectives, which may change his action. Instead of $a(\zeta)$, he chooses action $a^L(\zeta)$, where L signifies the presence of a "latent", unobserved experimenter demand influence. The influence could increase or decrease a: $a^L(\zeta) \gtreqless a(\zeta)$. We define the *latent demand effect* in environment ζ as the difference $a^L(\zeta) - a(\zeta)$.

While nonzero latent demand automatically biases estimates of mean actions, it does not necessarily bias estimates of treatment effects. To see this, note that the observed treatment effect can be decomposed as follows:

$$a^L(\zeta_1) - a^L(\zeta_0) = \underbrace{a(\zeta_1) - a(\zeta_0)}_{\text{Effect of interest}} + \underbrace{[a^L(\zeta_1) - a(\zeta_1)]}_{\text{Latent demand in } \zeta_1} - \underbrace{[a^L(\zeta_0) - a(\zeta_0)]}_{\text{Latent demand in } \zeta_0} \quad (2.1)$$

The first term on the right-hand side is the treatment effect of interest. The second and third capture the potential bias due to experimenter demand. If both demand effects are equal they cancel and the treatment effect is identified, but they may not cancel, either because the participant's inference or his response to a given inference varies with ζ. The usual logic of a randomised experiment is to ensure that variation in treatment is orthogonal to potential confounds, but as demand effects may be driven by the treatment

[9] In some experiments, the experimenter essentially fills the role of a real-world authority figure. For example part of the real-world response to incentives might include a response to perceived demand from an employer. For a researcher interested in the total effect of incentives, perceived demand may actually be part of the environment of interest, ζ, rather than a confound.

itself, randomisation does not guard against bias.

Consider two variants on the Dictator game, in which a participant is told to choose what fraction of \$10 to give to a recipient. In variant 0, he is told that the recipient is aware that the choice is taking place, while in variant 1 they are unaware (for instance, the money will just be added to a show-up fee). Absent any motive for pleasing the experimenter, the participant would prefer to give \$4, so the true treatment effect is $a(\zeta_1) - a(\zeta_0) = \$0$. However, in variant 0 he infers that the experimenter wants him to be generous, so he gives $\$5$, while in variant 1 he infers that the experimenter wants him to be selfish, so he gives zero. The experimenter fails to measure true preferences in either case, and identifies a treatment effect that is in reality a demand effect.

2.2.1 Demand treatments

We now assume that the experimenter has at her disposal a particular kind of treatment manipulation which we call a *demand treatment*. Negative demand treatments deliberately signal a demand that the decision-maker decrease his action, inducing $a^-(\zeta)$, while positive demand treatments demand an increase and induce $a^+(\zeta)$. Our first substantive assumption is a basic monotonicity condition:

Assumption 1 (*Monotonicity*)**.** $a^-(\zeta) \leq a^L(\zeta) \leq a^+(\zeta)$.

Assumption 1 requires that demanding an increased action does not decrease it, and vice versa. It has a natural connection to the monotonicity condition in the estimation of local average treatment effects (Angrist and

Imbens, 1994): the assumption rules out "defier" behaviour whereby participants do the opposite of what is demanded.

Our main assumption amounts to assuming that the demand treatments can bound the natural action of interest:

Assumption 2 (*Bounding*)**.** $a^-(\zeta) \leq a(\zeta) \leq a^+(\zeta)$.

It implies bounds for natural actions (2.2) and treatment effects (2.3):

$$a(\zeta) \in [a^-(\zeta), a^+(\zeta)] \tag{2.2}$$

$$a(\zeta_1) - a(\zeta_0) \in [a^-(\zeta_1) - a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)] \tag{2.3}$$

For some purposes we may wish to be able to make comparative statements about demand in different environments. Although the latent demand effect is unobservable, the sensitivity of behaviour to demand treatments may be informative about it. First, we define what we mean by "sensitivity".

Definition 1 (Sensitivity)**.** *Sensitivity is the difference in actions under positive and negative demand treatments:* $S(\zeta) = a^+(\zeta) - a^-(\zeta)$.

Remark 1. *In addition to bounding the natural action, assumptions 1 and 2 jointly imply that sensitivity* $S(\tau)$ *provides an upper bound on the magnitude of the latent demand effect:* $S(\zeta) \geq \left|a^L(\zeta) - a(\zeta)\right|$.

This fact enables us to use sensitivity $S(\zeta)$ to make statements of *comparative ignorance*, in the sense that if $S(\zeta_1) > S(\zeta_0)$ there is more scope for large latent demand effects under ζ_1 than ζ_0. But it could nevertheless be that the true latent demand effect is larger under ζ_0. Our third assumption, Monotone Sensitivity, allows us to make concrete claims about magnitudes.

Definition 2 (Comparison classes)**.** *A comparison class $Z^C \subseteq Z$ is a set of environments such that Monotone Sensitivity holds for all $z \in Z^C$.*

Assumption 3 (*Monotone Sensitivity*)**.** *$S(z)$ is strictly increasing in $\left|a^L(z) - a(z)\right|$ for all $z \in Z^C$.*

Monotone Sensitivity permits statements such as "latent demand is stronger for participant pool A than participant pool B" or "latent demand is stronger under incentive scheme A than incentive scheme B". We derive some comparison classes below using our Bayesian model.

2.2.2 Bayesian model

We now provide a simple foundation for our main assumptions, and derive conditions under which they will or will not hold. The environment ζ determines the mapping from actions $a \in \mathbb{R}$ into outcomes or distributions over outcomes. The decision-maker's payoff is $v(a, \zeta)$, where v captures the payoff structure (mapping from actions to outcomes) and preferences (mapping from outcomes to utility). We assume v is strictly concave and differentiable, so the natural action $a(\zeta)$ solves $v_1(a(\zeta), \zeta) = 0$.

Latent demand Demand enters preferences as follows. Upon observing ζ, the decision-maker makes an inference about the experimenter's objective, $h \in \{-1, 1\}$. If $h = -1$, he believes the experimenter benefits from him taking low actions, while if $h = 1$ he believes she benefits from high actions. He has a preference, ϕ, for pleasing the experimenter, which we allow to depend upon ζ.[10] We remain agnostic about *why* the participant wishes

[10]We have in mind that ϕ might depend on the identity of the experimenter (e.g. a firm versus a researcher) or decision-maker (e.g. women might have different attitudes than

to please the experimenter; possible motives include altruism, a motive to conform, or a belief that he will ultimately be rewarded for doing so.

We assume utility takes the following separable form:

$$U(a, \zeta) = v(a, \zeta) + a\phi(\zeta)E[h|\zeta]. \tag{2.4}$$

The optimal action $a^L(\zeta)$ thus solves:

$$v_1(a^L(\zeta), \zeta) + \phi E[h|\zeta] = 0 \tag{2.5}$$

so $a^L(\zeta) = a(\zeta) \Leftrightarrow \phi E[h|\zeta] = 0$. There is therefore no demand confound if either the decision-maker assigns equal likelihood to the preferred action being high or low ($E[h|\zeta] = 0$), or he does not care about the experimenter's objectives ($\phi = 0$) (these would be expected in a "natural field experiment," where the participant is unaware of the experiment). We assume the decision-maker's mean prior over h is $E[h] = 0$, so in the absence of any new information about h he chooses $a(\zeta)$. The relation between actions and beliefs is captured by $da^L(\zeta)/dE[h|\zeta] = -\phi/v_{11}(a, \zeta)$, which has the same sign as ϕ. Actions are monotone in beliefs.

We model learning as follows. The environment ζ includes a signal $h^L(\zeta) \in \{-1, 1\}$ which the decision-maker believes is a sufficient statistic, i.e. $E[h|h^L(\zeta), \zeta] = E[h|h^L(\zeta)]$. He believes that with probability $p^L(\zeta)$, the signal is correct ($h^L = h$), and with probability $1 - p^L(\zeta)$ it is pure noise ($h^L = \epsilon$, where ϵ equals -1 or 1 with equal probability). We impose that $p^L(\zeta) \in [0, 1)$. It is

men). ϕ might also vary with other features such as the salience of the benefit to the experimenter, or how important the participant believes his actions are for achieving the experimenter's objectives.

straightforward to see that:

$$E[h|h^L(\zeta)] = h^L(\zeta)p^L(\zeta) \tag{2.6}$$

The decision-maker's belief depends on ζ in two ways. First, via the sign of $h^L(\zeta)$, i.e. whether he believes that the experimenter wants a high or low action, which determines the *direction* of the latent demand effect. Second, via $p^L(\zeta)$, i.e. the perceived informativeness of the signal, which affects the *magnitude* of the latent demand effect.

Demand treatments We assume that the experimenter can choose a "demand treatment" signal $h^T \in \{-1, 1, \emptyset\}$. $h^T = \emptyset$ corresponds to the usual case in which no demand treatment is used, while $h^T = 1$ and $h^T = -1$ correspond to positive and negative demand treatments. These signals provide information about h so as to direct the decision-maker's beliefs. We assume that if $h^T = \emptyset$ the decision-maker does not update his belief about h (for example because their prior is that demand treatments are never used). This assumption is reasonable as (at present) demand treatments are rarely used in experiments. We maintain throughout that ζ (and hence $v(a, \zeta)$, $h^L(\zeta)$, $p^L(\zeta)$, and $\phi(\zeta)$) does not depend on the demand treatment, i.e. receiving a demand treatment does not change the decision-maker's interpretation of the maintained experimental environment or their motive for pleasing the experimenter. Instead the demand treatment is interpreted purely as informative about the direction of the experimenter's objective.[11]

[11]Formally, we assume that $\zeta(h^T) = \zeta$, $\forall \zeta$. This assumption will be stronger for some demand treatments and environments than others, and is an important consideration in the selection of appropriate demand treatments. If it does not hold then Bounding might fail because the demand treatments alter the natural action itself: $a(\zeta(\emptyset)) \notin [a(\zeta(-1)), a(\zeta(1))]$. In Appendix 2.B.5 we extend the model to allow ϕ to depend on h^T and show that the Bounding condition remains unchanged.

The decision-maker believes that h^T is informative about h: with probability p^T, h^T equals h, and with probability $1 - p^T$ it equals η, which takes values -1 and 1 with equal probability. η and ϵ are believed to be independent (we revisit this assumption in Appendix 2.B.6). The Bayesian posterior is:

$$E[h|h^T, h^L(\zeta)] = \frac{h^L(\zeta)p^L(\zeta) + h^T p^T}{1 + h^L(\zeta)p^L(\zeta)h^T p^T} \tag{2.7}$$

Thus, if $h^L(\zeta) = h^T$, the demand treatment reinforces the participant's belief, while if the signals have opposite signs they offset one another.

Assumptions We now use the model to provide foundations for our main assumptions described in Section 2.2.1. Derivations can be found in Appendix 2.B.

First, Assumption 1 (Monotonicity) states that a positive demand treatment increases the action (relative to no demand treatment) and the negative demand treatment decreases it. It is straightforward to see that except for the trivial case $p^T = 0$, these conditions are satisfied if and only if $\phi \geq 0$, i.e. a weak preference for pleasing the experimenter.

Proposition 1. *Monotonicity holds for all p^T if and only if $\phi \geq 0$.*

Second, Assumption 2 (Bounding) states that the demand treatments provide bounds on the true action. In the Bayesian model, given $\phi \geq 0$ (Monotonicity), the action is larger or smaller than $a(\zeta)$ when $\phi E[h|h^T, h^L] \geq 0$ or $\phi E[h|h^T, h^L] \leq 0$ respectively. Intuitively, whatever the latent demand effect, the demand treatment that opposes it must be informative enough to reverse the sign of beliefs. It is clear from inspection of (2.7) that this simply requires the demand treatments to be "more informative" than latent

demand, $p^T \geq p^L(\zeta)$.

Proposition 2. *Given $\phi \geq 0$, Bounding holds if and only if $p^T \geq p^L(\zeta)$.*

Finally, Assumption 3 (Monotone Sensitivity) states that within a comparison class Z^C of environments, differences in sensitivity are informative about differences in underlying latent demand. Latent demand and sensitivity can vary for multiple reasons, so there is no simple condition that guarantees when this assumption will and will not hold. In Appendix 2.B.3 we work out some important cases. First, we show that Monotone Sensitivity holds when variation in demand effects is driven by differences in the strength of preference for pleasing the experimenter, ϕ. Second, we analyze Monotone Sensitivity when variation in demand effects is driven by differences in the payoff function, v, deriving specific conditions when v is additively or multiplicatively separable and providing examples such as variation in incentives. Third, we show that Monotone Sensitivity holds in a model of inattention to experimenter demand. Finally, we show that Monotone Sensitivity does *not* hold in general when environments differ in the beliefs they induce ($E[h|h^L(\zeta)]$). We use these findings when interpreting heterogeneous responses to demand treatments in section 2.5.4.

2.2.3 "Weak" and "strong" demand treatments

There are many different ways to signal a desire for high or low actions. How should the experimenter choose? The model gives us a way to answer this question. The width of the bounds $[a^-(\zeta), a^+(\zeta)]$ is increasing in p^T. Therefore the tightest bounds, subject to satisfying Bounding ($p^T \geq p^L(\zeta)$), are obtained when $p^T = p^L(\zeta)$. In other words, we want the "least infor-

mative" demand treatment possible, subject to being "informative enough" for Bounding.[12] We want to choose demand treatments that are likely to be "stronger" or more informative than any latent demand in the study of interest, while avoiding excessively strong signals that lead to uninformative bounds.

In our empirical applications we employ two types of demand treatments, described in more detail below. Our "weak" manipulations explicitly signal what we expect participants to do; we believe these are already more informative than likely latent demand in typical experiments. Our "strong" manipulations go further, telling participants which action will "do us a favour". These lead to more conservative bounds, and may be useful for applications where researchers are especially concerned about demand effects. They also play a role in more structural applications, described in Sections 2.4.4 and 2.4.5.

2.2.4 Heterogeneity and Defiers

The approach naturally extends to the case where participants are heterogeneous and the experimenter is interested in average behavior or average treatment effects. If Monotonicity and Bounding hold for all agents individually, then they also hold for average actions, so we can simply reinterpret a, a^L, a^+ and a^- as representing average behaviors and our approach remains valid.

An important dimension of heterogeneity is in ϕ, the preference for pleasing

[12]This gives a novel reason why deception in experiments can be problematic. If the demand treatment is regarded as uninformative because participants are used to second-guessing what experimenters are really after, then the bounding exercise is invalidated. We thank an anonymous referee for this observation.

the experimenter. Monotonicity requires a weak positive preference, $\phi \geq 0$. "Defiers" with $\phi < 0$ prefer to go against the experimenter's wishes. Bounding fails for these individuals, because $a^- > a^+$. We show in Appendix 2.B.4 that the method is able to tolerate some defier behavior, but too much will lead to failures of Bounding. We give an example where Bounding is satisfied provided the average participant is a complier. In general, for defier behaviour to be "small enough" the joint distribution of preferences and beliefs must be such that that the response by the compliers outweighs that of the defiers.

2.3 Sample and experimental design

We conducted seven experiments in total to demonstrate our approach and to provide estimates of demand sensitivity on a wide range of standard experimental tasks. Our respondents complete one of the following tasks: a dictator game (Kahneman et al., 1986); a risky investment game (Gneezy and Potters, 1997), without or with ambiguity; a convex time budget task (Andreoni and Sprenger, 2012); a trust game (first or second mover, Berg et al., 1995); an ultimatum game (first or second mover, Güth et al., 1982); a lying game (Fischbacher and Föllmi-Heusi, 2013); and a real effort task with or without performance pay (DellaVigna and Pope, 2017, DellaVigna and Pope, 2018). We conduct all of our experiments online, primarily because the large number of treatments would be infeasible to implement in the laboratory. We designed the experiments to maximize comparability. For all experiments except the effort task, the action spaces are similar (they can be expressed as real numbers from 0 to 1); we pay the same show-up fee; recruit from the same participant pools; use the same mode of collec-

tion (online); the same response mode (sliders); and keep stakes as similar as possible.[13]

We employ two phrasings for our demand treatments. Our "weak" treatments explicitly tell participants that we expect high or low actions. For example, in the investment game, participants were told at the end of their instructions that "We expect that participants who are shown these instructions will invest more/less in the project than they normally would".[14] The strong treatments go further, telling participants that they will "do us a favour" by taking a higher or lower action. For example, in the dictator game, participants in the positive demand condition were told "You will do us a favour if you give more/less to the other participant than you normally would". We keep the phrasing of the demand treatments as homogeneous as possible across tasks. In the two-player games we do not provide information about demand treatments shown to the other player, but our approach could be extended to create common knowledge about demand. Table 2.1 summarizes the design features of each experiment, and Table 2.2 provides design details, parameters, and the exact wording of the demand treatments for each task. Figure A.1 gives an example from the experimental interface. Full experimental instructions can be found in the pre-analysis plans available at: `https://www.socialscienceregistry.org/trials/1248`.

[13]For the effort task, we replicated the design employed in DellaVigna and Pope (2017) and DellaVigna and Pope (2018). The primary differences with our other tasks are a higher show-up fee and a different response mode (effort).

[14]It is not completely straightforward to design demand treatments that report the experimental hypothesis, because if the experimenter truly hypothesizes that the action will be high in one treatment, telling participants she expects it to be low could be considered deceptive. By referring to "participants who are shown these instructions" (which include the demand treatment) we avoid this issue, because it is indeed true that we expect high actions from participants in the positive demand treatment group and low actions in the negative demand treatment group.

Table 2.1: Overview of Experiments

Experiment	Sample	Tasks	Demand Treatments	Real or Hypothetical
Experiment 1 (May 18, 2016 - May 30, 2016)	MTurk (N=4,479)	Dictator Game, Investment Game and Convex Time Budgets	Strong positive demand, strong negative demand and no-demand treatment	Both real stakes and hypothetical choices
Experiment 2 (July 5, 2016 - July 25, 2016)	MTurk (N=2,950)	Dictator Game and Investment Game	Weak positive demand, weak negative demand and no-demand treatment	Both real stakes and hypothetical choices
Experiment 3 (Aug 26, 2016 - Aug 27, 2016)	MTurk (N=1,691)	Effort experiment with 1 cent bonus and Effort experiment with no bonus. Also effort experiment with 4 cent bonus (no demand treatments were applied to this group).	Strong positive, strong negative and no-demand treatment	Real stakes (real effort experiment)
Experiment 4 (Aug 18, 2016 - Sep 1, 2016 7)	Research Now Representative online Panel (N= 2,933)	Dictator Game and Investment Game	Strong positive demand, strong negative demand, weak positive demand and weak negative demand and no-demand treatment	Real stakes
Experiment 5 (Sep, 12, 2016 - Sep 20, 2016)	MTurk (N= 5,045)	Trust game (first and second mover), Ultimatum game (first and second mover), Lying game, Ambiguous Investment Game and Convex Time Budgets	Strong positive demand, strong negative demand, weak positive demand and weak negative demand	Real stakes
Experiment 6 (Sep, 19, 2016 - Sep 20, 2016)	MTurk (N=769)	Effort experiment with 1 cent bonus, Effort experiment with no bonus	Weak positive demand and weak negative demand	Real stakes (real effort experiment)
Experiment 7 (May 18, 2017 - May 20, 2017)	MTurk (N=999)	Dictator Game and Investment Game	Within design: Task 1: no demand treatment; task 2: strong positive demand or strong negative demand	Real stakes

Notes: This table summarizes the key design features of each of the experiments. In experiment 5 for convex time budgets we only employ the weak demand treatments, while for all other tasks in experiment 5 we employ both strong and weak demand treatments.

Table 2.2: Details of Experimental Tasks

Task	Description	Sample	Show up fee	Action space	Payoffs	Strong Demand Treatment: "You will do us a favour if you…"	Weak Demand Treatment: "We expect that participants who are shown these instructions will…"
Dictator Game	Choose to split money between yourself and another participant.	Strong: MTurk between (N=1,503); MTurk within (998); Rep. Sample (N=899). Weak: MTurk between (N=1,477); Rep. Sample (N=865).	$.25	$a \in [0,1]$ is fraction of $1 sent to the other participant.	Amount not sent $(1-a)$.	"…give more (less) to the other participant than you normally would".	"…give more (less) to the other participant than they normally would".
Investment Game	Choose how much to invest in a risky project.	Strong: MTurk between (N=1,492); MTurk within (1000); Rep Sample (N=902). Weak: MTurk between (N=1,473); Rep. Sample (N=883).	$.25	$a \in [0,1]$ is fraction of $1 invested	$\$(1-a)+3a$ with probability 0.4, $\$(1-a)$ otherwise.	"…invest more (less) than you normally would".	"…invest more (less) than they normally would".
Ambiguous Investment Game	Choose to how much to invest in an uncertain project.	Strong: MTurk (N= 404). Weak: MTurk (N= 390).	$.25	$a \in [0,1]$ is fraction of $1 invested	$\$(1-a)+3a$ with unstated probability, $\$(1-a)$ otherwise.	"…invest more (less) than you normally would".	"…invest more (less) than they normally would".
Convex Time Budgets	Choose between receiving money today vs. money in seven days.	Strong: MTurk (N=1,484); Rep. Sample (N=899). Weak: MTurk (N=422).	$.25	$a \in [0,1.2]$ is the amount to be received in 7 days	$\$(1-a)/1.2$ is received within 24 hours, and $\$a$ is received in 7 days.	"…choose to receive more (less) in seven days than you normally would".	"…choose to receive more (less) in seven days than they normally would".
Effort: No bonus	Alternately press the a and b button without receiving any bonus.	Strong: MTurk (N=731). Weak: MTurk (N=388).	$1	$a \in [0,4000]$ is number of a-b button presses	No payoffs beyond show-up fee	"…work harder (less hard) than you normally would".	"…work harder (less hard) than they normally would".
Effort: 1-cent bonus	Alternately press the a and b button, receiving 1 cent per 100 points.	Strong: MTurk (N=714). Weak: MTurk (N=381).	$1	$a \in [0,4000]$ is number of a-b button presses	1 cent per 100 button presses.	"…work harder (less hard) than you normally would".	"…work harder (less hard) than they normally would".
Trust Game 1st mover	Choose to send an amount of money to the other player.	Strong: MTurk (N=382). Weak: MTurk (N=352).	$.25	$a \in [0,.2,.4,.6,.8,1]$ is fraction of $1 sent	$\$2a$ is sent to second mover, who decides how much to send back. $\$(1-a)$ not sent is kept with certainty.	"…send more (less) to the other participant than you normally would".	"…send more (less) to the other participant than they normally would".
Trust Game 2nd mover	Choose to send back some money to the other player. (Strategy method)	Strong: MTurk (N=371). Weak: MTurk (N=346).	$.25	$a \in [0,1.2]$ is amount returned, averaged over each possible nonzero amount received.	Amount not sent back.	"…send back more (less) to the other participant than you normally would".	"…send back more (less) to the other participant than they normally would".
Ultimatum Game 1st mover	Offer a split to the other player.	Strong: MTurk (N=409). Weak: MTurk (N=360).	$.25	$a \in [0,1]$ is offer to the other player	$1-a$ if the offer is accepted, 0 if it is rejected.	"…offer more (less) to the other participant than you normally would".	"…offer more (less) to the other participant than they normally would".
Ultimatum Game 2nd mover	Specify the smallest offer you would accept.	Strong: MTurk (N=421). Weak: MTurk (N=411).	$.25	$a \in [0,1]$ is min. acceptable offer: reject all offers below this amount.	Amount received if it exceeds a, otherwise zero.	"…require a higher (lower) minimum amount than you normally would".	"…require a higher (lower) minimum amount than they normally would".
Lying	Report the number of "Heads" in 10 coinflips.	Strong: MTurk (N=365). Weak: MTurk (N=412).	$.25	$a \in [0,1,\ldots,10]$ is number of heads.	10 cents per "Heads" reported: $\$0.1a$.	"…report more (fewer) heads than you normally would".	"…report more (fewer) heads than they normally would".

2.3.1 Participant populations

We conducted six experiments with approximately 16,000 participants (or "workers") on Amazon Mechanical Turk (MTurk) (Experiments 1–3 and 5–7), and one experiment with around 3,000 participants using an online panel sample representative of the US population in terms of region, age, income, and gender (Experiment 4). MTurk is an online labor marketplace that is frequently used by researchers for surveys and experiments. It is attractive because it offers a large and diverse pool of workers. There is some evidence that MTurk workers are more attentive to instructions than college students (Hauser and Schwarz, 2016). To participate in our MTurk experiments, workers had to live in the U.S, have an overall approval rating of more than 95 percent, and have completed more than 500 tasks on MTurk, fairly standard parameters in research on MTurk.[15]

Most workers on MTurk are experienced in taking surveys, which might affect the external validity of our results. We used the representative sample, whose participants are less experienced with social science experiments, to replicate a subset of our findings. The sample is maintained by a market research company, *Research Now*.

[15]We excluded prior participants when recruiting for experiments 2 and 3. Technically this is achieved by applying a "qualification" flag to the MTurk accounts of prior participants, which can then be used to prevent them seeing or accepting new MTurk tasks posted by us. At the time of running experiments 5 and 6, we had essentially exhausted the active participant pool, and to avoid undue delays in recruitment we therefore allowed prior participants to take part. Around 36 percent of the respondents in these experiments had not participated before. In experiment 7, which was conducted some time later, we did exclude prior participants, but a server communication error meant that not all accounts received the qualification flag and as a result some prior participants did take part. 70 percent of the respondents in this experiment had not participated before. Our results are virtually unchanged by the dropping of participants who completed more than one of our experiments; results are available upon request.

2.3.2 Pre-analysis plans

Our experiments were conducted in a sequence, between May 2016 and May 2017. Each is described in a pre-analysis plan (PAP) posted online prior to launch.[16] The sequence is laid out in Table 2.1. For each experiment, the PAP details the data to be collected, treatment variables, experimental instructions, and how we planned to analyze that experiment's data.

However, presenting the data experiment-by-experiment is repetitious. Therefore, for brevity and clarity of exposition, in the paper we pool the data and analyze all tasks side-by-side for our weak and strong demand treatments separately (this structure was described in pre-analysis plan 5). Our main analysis uses data from MTurk respondents with real stakes, which we have for all eleven tasks studied. In the analysis of heterogeneity we introduce hypothetical choice data from MTurk and the representative panel, which were collected for a subset of tasks. When averaging across tasks we weight observations to give equal weight to each task.

Other than this pooling across experiments, our analysis closely follows what was pre-specified.[17] All pre-specified analyses are available in Appendix C on the following link: `https://www.dropbox.com/s/i5vkznc3crnh7r9/deQuidt_Haushofer_Roth_demand_final.pdf?dl=0`.

[16] The pre-analysis plans were posted on the Social Science Registry and can be found here: https://www.socialscienceregistry.org/trials/1248.

[17] In some experiments we proposed to standardize responses based on average choices in the no-demand condition. Because we did not collect no-demand data for all tasks, for consistency we always standardize based on the negative demand treatment group (a simple and inconsequential linear transformation). For our real-effort tasks, which were based on DellaVigna and Pope (2018), we pre-specified that we would apply their exclusion criteria to the analysis dataset (excluding participants that take more than 30 minutes, take the task more than once, score zero or more than 4,000 points, or have invalid MTurk IDs). In our other experiments we did not pre-specify exclusions, but for consistency we also drop participants who submitted multiple responses (less than 0.5 percent). This is inconsequential for the results.

2.3.3 Summary statistics

Tables A.13 to A.19 in the Appendix present the pre-specified balance tables for all of the experiments. Tables A.20 to A.27 provide summary statistics on our respondents. Table A.24 highlights that respondents from the online panel are representative of the US population by gender, income, age, and region, and other observables. Attrition was low, below 2 percent on average, and did not differ across demand treatment arms (Tables A.28 and A.29).

2.4 Applying the method

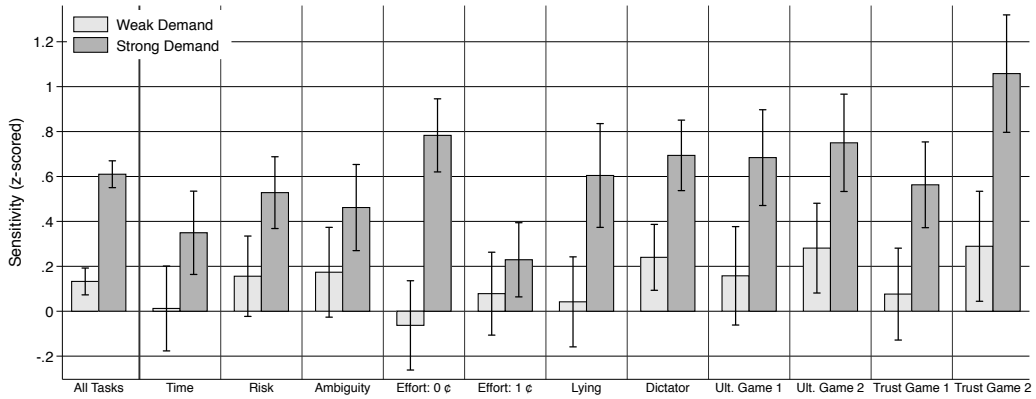
2.4.1 Bounding natural actions

In this section we provide bounds on natural actions estimated using our weak and strong demand treatments. For a subset of tasks we also measured behaviour with no demand treatment, and describe these results in Section 2.5.1 where we discuss Monotonicity. Our objects of interest here are mean behaviour in the positive ($a^+(\zeta)$) and negative ($a^-(\zeta)$) demand conditions.

Panel A of Table 2.3 and Figure 2.2 show mean actions by task and demand treatment for incentivised MTurk respondents with weak treatments. Panel B of Table 2.3 and Figure 2.1 display sensitivities ($a^+(\zeta) - a^-(\zeta)$), in both raw and z-scored units. Sensitivity is modest, averaging around 0.13 standard deviations, and frequently not significantly different from zero. The strongest responses (between 0.2–0.3 standard deviations) were observed

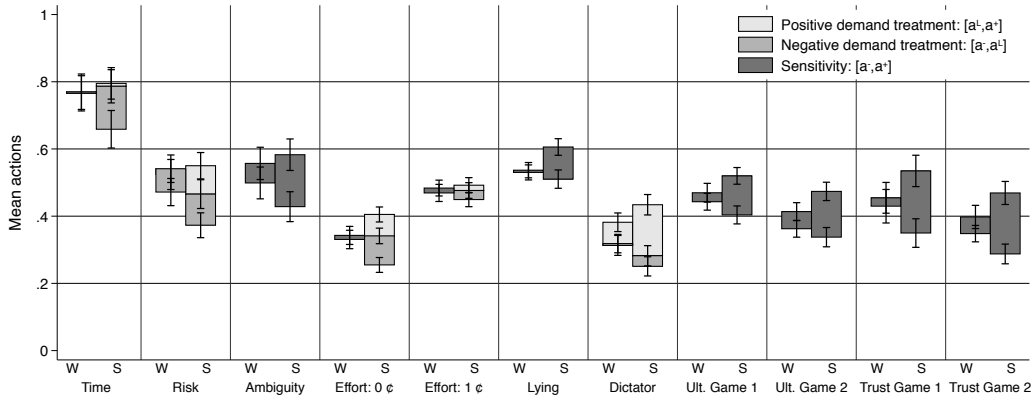
for the dictator game, the ultimatum game second mover, and the trust game second mover. As we have argued, the weak manipulations seem likely to satisfy bounding for typical applications, so these results give cause for optimism. Panel A of Table 2.4 and Figure 2.2 show mean actions in the different demand treatment arms employing strong treatments. Panel B of Table 2.4 and Figure 2.1 display sensitivities. behaviour is responsive to our strong demand treatments, and sensitivity is significantly different from zero in all tasks, averaging around 0.6 standard deviations. Sensitivity is particularly high in the dictator game, for second movers in the trust and ultimatum games, and for unincentivised effort. These manipulations are significantly stronger than likely implicit signals in most experiments or surveys, so providing quite conservative upper bounds on typical demand biases. However, they do demonstrate that participants are motivated to respond to signals about the researcher's goals, and that responses can be significant when those signals are strong. Thus the attention researchers pay to potential demand effects at the study design stage is well justified.

Figure 2.1: Sensitivity to demand treatments, z-scored



Notes: This figure uses data from incentivised MTurk respondents with weak and strong demand treatments. It presents the z-scored sensitivity of behaviour to our demand treatments, i.e. the normalised difference in behaviour between the positive and negative demand conditions. Error bars indicate 95 percent confidence intervals.

Figure 2.2: Bounding natural actions



Notes: This figure uses data from incentivised MTurk respondents with weak (W) and strong (S) demand treatments. It displays mean responses by task and demand treatment. Upper (lower) points correspond to positive (negative) demand treatments (a^+ and a^-), intermediate points to "no demand" treatments (a^L, not collected for all tasks). Lighter shaded sections indicate the response to positive and negative demand treatments separately, dark shaded sections indicate sensitivity when a^L was not measured. Error bars indicate 95 percent confidence intervals.

Table 2.3: Response to weak demand treatments, all incentivised tasks

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.770	0.524	0.557	0.331	0.484	0.537	0.382	0.470	0.413	0.455	0.398
	(0.027)	(0.023)	(0.024)	(0.014)	(0.012)	(0.012)	(0.014)	(0.014)	(0.014)	(0.023)	(0.017)
No demand		0.541					0.313				
		(0.021)					(0.015)				
Negative demand	0.766	0.472	0.499	0.343	0.469	0.530	0.318	0.443	0.362	0.430	0.348
	(0.027)	(0.021)	(0.024)	(0.014)	(0.013)	(0.011)	(0.014)	(0.013)	(0.013)	(0.025)	(0.012)
Panel B: Sensitivity (positive - negative)											
Raw data	0.005	0.052	0.058	-0.012	0.015	0.007	0.063	0.027	0.051	0.025	0.050
	(0.038)	(0.031)	(0.034)	(0.019)	(0.018)	(0.016)	(0.020)	(0.019)	(0.019)	(0.034)	(0.021)
Z-score	0.012	0.156	0.174	-0.063	0.078	0.042	0.240	0.158	0.281	0.076	0.289
	(0.096)	(0.091)	(0.102)	(0.101)	(0.094)	(0.102)	(0.075)	(0.112)	(0.102)	(0.104)	(0.125)
		[0.096]					[0.002]				
Panel C: Monotonicity											
Positive - Neutral (z-score)		-0.051					0.261				
		(0.092)					(0.078)				
		[0.237]					[0.002]				
Negative - Neutral (z-score)		-0.207					0.021				
		(0.087)					(0.078)				
		[0.056]					[0.357]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Notes: This table uses data from incentivised MTurk respondents with weak demand treatments. Panel A displays mean actions with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of behaviour to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task when testing the Monotonicity assumption.

Table 2.4: Response to strong demand treatments, all incentivised tasks

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.795	0.550	0.583	0.405	0.492	0.606	0.434	0.520	0.474	0.535	0.469
	(0.024)	(0.020)	(0.024)	(0.011)	(0.011)	(0.013)	(0.015)	(0.013)	(0.014)	(0.024)	(0.017)
No demand	0.786	0.466		0.341	0.476		0.282				
	(0.025)	(0.022)		(0.012)	(0.012)		(0.015)				
Negative demand	0.659	0.373	0.428	0.255	0.449	0.510	0.251	0.404	0.337	0.350	0.288
	(0.028)	(0.019)	(0.023)	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)	(0.015)	(0.022)	(0.015)
Panel B: Sensitivity (positive - negative)											
Raw data	0.137	0.177	0.155	0.150	0.043	0.096	0.183	0.116	0.136	0.185	0.181
	(0.037)	(0.027)	(0.033)	(0.016)	(0.016)	(0.019)	(0.021)	(0.018)	(0.020)	(0.032)	(0.023)
Z-score	0.349	0.528	0.462	0.783	0.229	0.604	0.694	0.684	0.750	0.563	1.058
	(0.095)	(0.082)	(0.098)	(0.083)	(0.084)	(0.118)	(0.080)	(0.109)	(0.111)	(0.097)	(0.133)
	[0.001]	[0.001]		[0.001]	[0.020]		[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)	0.022	0.252		0.333	0.084		0.574				
	(0.088)	(0.088)		(0.085)	(0.088)		(0.082)				
	[0.363]	[0.001]		[0.001]	[0.159]		[0.001]				
Negative - Neutral (z-score)	-0.327	-0.276		-0.450	-0.145		-0.120				
	(0.097)	(0.086)		(0.085)	(0.086)		(0.080)				
	[0.001]	[0.001]		[0.001]	[0.101]		[0.046]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Notes: This table uses data from incentivised MTurk respondents with strong demand treatments. Panel A displays mean actions with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of behaviour to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task when testing the Monotonicity assumption.

2.4.2 Bounding treatment effects

Our real effort experiments replicate treatments from DellaVigna and Pope (2018). Participants alternately pressed the "a" and "b" keyboard buttons for 10 minutes, earning one point per pair. One group were told that their score "will not affect [their] payment," while a second group received one cent per 100 points. By combining the bounds estimated for each incentive treatment we can construct bounds on the treatment effect of performance pay on effort provision.[18]

Table 2.5 displays the conventional treatment effect ($a^L(1) - a^L(0)$, where "1" and "0" correspond to the reward per 100 points), the upper bound of the treatment effect ($a^+(1) - a^-(0)$), and the lower bound ($a^-(1) - a^+(0)$). In words, the lower bound on the treatment effect is given by comparing participants who received performance pay, coupled with a negative demand treatment, to participants who received no performance pay, coupled with a positive demand treatment. We first show the bounds generated using our weak treatments, which are quite tight, ranging from 0.67 to 0.75 standardised units.[19] The width of these bounds corresponds to only 11 percent of the estimated treatment effect (or 0.07 standard deviations), suggesting

[18]Our pre-analysis plans did not explicitly describe the bounding of treatment effects, but it is an immediate extension of the approach to bounding actions.

[19]In constructing the bounds using our weak treatments we note that the average effort in the no-incentive condition was actually slightly higher for those receiving negative demand than those receiving positive demand, i.e. we observe a small monotonicity failure ($a^+(0) < a^-(0)$). When sensitivity is low, such outcomes can easily arise due to sampling variation; both values here are statistically indistinguishable. In such cases, the procedure we propose in this section could lead to bounds on the treatment effect with negative width. A conservative approach, which we follow, is to first "iron" the bounds on the no-incentive condition, by averaging them. Formally, one can compute $a^+_{iron}(\zeta) = \max\{a^+(\zeta), 0.5[a^+(\zeta) + a^-(\zeta)]\}$ and $a^-_{iron}(\zeta) = \min\{a^-(\zeta), 0.5[a^+(\zeta) + a^-(\zeta)]\}$, and then use these values when computing the bounds on the treatment effect, which become $a^+(1) - a^-_{iron}(0)$, and $a^-(1) - a^+_{iron}(0)$. Because in this case $a^+_{iron}(0) = a^-_{iron}(0)$, the width of the weak bounds on the treatment effect is simply equal to $a^+(1) - a^-(0)$.

a limited role for experimenter demand in explaining the effort response to incentives. Naturally, the bounds created using the more conservative strong treatments are much wider, ranging from 0.23 to 1.21 standard deviations. Even these conservative bounds support the qualitative finding that effort responds to incentives.

Table 2.5: Bounding Treatment Effects

	Conventional	Weak Bounds		Strong Bounds	
	Treatment Effect	Lower	Upper	Lower	Upper
Count	540.720	530.001	588.270	177.421	948.978
	(66.763)	(64.532)	(61.499)	(62.379)	(64.148)
Count (z-scored)	0.686	0.673	0.747	0.225	1.205
	(0.085)	(0.082)	(0.078)	(0.079)	(0.081)

Notes: This table uses data from the real effort experiments with weak and strong demand treatments (experiments 3 and 6). Column 1 shows conventional treatment effect estimates. Columns 2 to 6 show lower and upper bounds estimated using weak and strong treatments. We apply the "ironing" procedure described in section 2.4.2 when constructing the weak estimates. Robust standard errors in parentheses. "Count" is the raw score from the experiment, Count (z-scored) is standardised using the negative demand condition, pooled across incentive treatment arms.

2.4.3 Confidence intervals

It is possible to compute confidence intervals for (a) the bounds themselves, and (b) the parameters contained by those bounds (a natural action or treatment effect), following Imbens and Manski (2004) (see Appendix 2.B.7 for details). The latter can be thought of as "demand-robust" confidence intervals, combining conventional parameter uncertainty due to sampling error with the additional uncertainty about possible demand effects. Uncertainty due to sampling error can be reduced in the usual way by increasing sample size (specifically, in the demand treatment arms), while uncertainty due to demand is reduced by selecting minimally informative demand treatments,

subject to Bounding (see section 2.2.3). Table A.3 presents confidence intervals computed from individual tasks using both the weak and strong demand treatments. Table A.4 presents confidence intervals on the bounds and treatment effect of the effect of incentive pay in the effort experiment. Zero lies outside these confidence intervals, providing statistical support for the finding that incentives increased effort.

2.4.4 Controlling for demand

The nonparametric bounding approach described above yields bounds on treatment effects, but researchers may be interested in point estimates that "control for" demand effects. Intuitively, one might apply same-signed demand treatments (positive-positive or negative-negative) to the treatment group and the control group, with the goal of harmonizing demand between treatments. In this section we describe how using this approach can eliminate bias if demand treatments are assumed to be fully informative ($p^T = 1$), and can reduce bias in other cases. Derivations are given in Appendix 2.B.8.[20]

We will assume throughout that Monotonicity holds strictly, i.e. $\phi > 0$ ($\phi = 0$ would imply no demand bias). The participant's usual first-order condition, with demand treatment h^T and optimal action $a^*(\zeta, h^T)$, is $v_1(a^*(\zeta, h^T), \zeta) + \phi(\zeta)E[h|h^T, h^L(\zeta)] = 0$. A first-order Taylor approximation around the natural action $a(\zeta)$ yields:

$$a^*(\zeta, h^T) \approx a(\zeta) + \Phi(\zeta)E[h|h^T, h^L(\zeta)] \tag{2.8}$$

[20]We thank the editor, Stefano DellaVigna, as well as an anonymous referee for suggesting this line of inquiry.

where $\Phi(\zeta) \equiv -\phi(\zeta)/v_{11}(a(\zeta), \zeta)$ is a slope term capturing the effect of beliefs on actions, which we term "responsiveness". Φ is positive as $v_{11} < 0$.

Assume two treatment groups, $\zeta \in \{0, 1\}$, with identical demand treatments $h^T \in \{-1, 1, \emptyset\}$, from which we estimate a treatment effect $a^*(1, h^T) - a^*(0, h^T)$. Its bias relative to the true effect can be decomposed as follows:

$$
\begin{aligned}
Bias &= [a^*(1, h^T) - a^*(0, h^T)] - [a(1) - a(0)] \\
&\approx \underbrace{\Phi(1)\left(E[h|h^T, h^L(1)] - E[h|h^T, h^L(0)]\right)}_{\text{Bias due to beliefs}} + \underbrace{(\Phi(1) - \Phi(0))\, E[h|h^T, h^L(0)]}_{\text{Bias due to "responsiveness"}}
\end{aligned}
$$

The first term captures differences in beliefs between the treatment and control environments, for example because they induce differences in latent demand. The second captures differences in behavioural responsiveness, given beliefs, for example because the treatment and control groups are at different locations on the cost of effort function.[21]

Fully informative demand treatments Importantly, in the special case where researchers are willing to assume that demand treatments are fully informative ($p^T = 1$), we can eliminate the bias due to beliefs: if h^T is fully informative, $E[h|h^T, h^L(1)] = E[h|h^T, h^L(0)] = 1$ or -1. We are left with the bias due to differences in responsiveness. We can then ask whether this bias is important, by testing for differences in sensitivity between treatment and

[21] In some settings it may be possible to sign the bias due to responsiveness. If demand treatments are applied, and bounding holds, the sign of $E[h|h^T, h^L(0)]$ is known and equal to the sign of h^T. Knowledge of the shape of v can then help us to sign $\Phi(1) - \Phi(0)$. For example in the real effort case, we expect responsiveness to decrease as effort increases, due to the curvature of the cost of effort function. That implies $\Phi(1) - \Phi(0) < 0$, in which case the bias due to responsiveness is negative when positive demand treatments are used.

control (an interaction effect):[22]

$$\underbrace{[a^*(1,1) - a^*(1,-1)]}_{\text{Sensitivity } (\zeta = 1)} - \underbrace{[a^*(0,1) - a^*(0,-1)]}_{\text{Sensitivity } (\zeta = 0)} \approx 2\left(\Phi(1) - \Phi(0)\right).$$

If this term is small, we can obtain a point estimate of the demand-free treatment effect by comparing behaviour on two same-signed demand treatments, essentially we are "controlling for" the influence of demand.

If sensitivity differs significantly between treatment and control, we can still approximate the treatment effect by averaging the estimates obtained with two positive and two negative demand treatments:

$$0.5\left([a^*(1,1) - a^*(0,1)] + [a^*(1,-1) - a^*(0,-1)]\right) \approx a(1) - a(0).$$

This approach is equivalent to estimating the treatment effect from the midpoints of the bounds for the treatment and control groups. It relies on the symmetry of the first-order Taylor approximation.

Less informative treatments Alternatively, researchers might wish to use same-signed weaker demand treatments to align beliefs among participants, without requiring $p^T = 1$. In general this will not eliminate bias entirely, but we can derive conditions under which the bias will be reduced. Since differences in responsiveness will no longer be testable we focus on the prospect of reducing the bias due to beliefs, which will be sufficient if variation in responsiveness between treatments is small.[23] We find that when the latent demand biases have opposite signs ($h^L(1) = -h^L(0)$, which is the typical

[22] Or, equivalently, testing whether the treatment effect estimate differs when two positive versus two negative demand treatments are used.

[23] In other words we ask when $\left|E[h|h^T, h^L(1)] - E[h|h^T, h^L(0)]\right| < \left|E[h|h^L(1)] - E[h|h^L(0)]\right|$, for $h^T \in \{-1, 1\}$.

scenario that concerns researchers) our Bounding assumption is sufficient for two same-signed demand treatments to reduce the bias due to beliefs. When the latent demand biases have the same sign ($h^L(1) = h^L(0)$), same-signed demand treatments that *reinforce* latent demand (i.e. $h^T = h^L(1)$) always reduce bias. Sufficiently strong opposite-signed treatments reduce bias, but Bounding is not enough to guarantee this.

In summary, the Bounding assumption covers all cases except where the demand effects in treatment and control agree with one another and disagree with the demand treatments used. To apply this approach, therefore, researchers may need to use judgment about the likely sign of demand effects in their experiment, or report a range of estimates.

Applications We apply the above-developed approaches to our effort experiment in Appendix Table A.1. For the strong demand treatments, where we have argued $p^T = 1$ is not an unreasonable assumption, we see large and statistically significant differences in sensitivity between the 0¢ and 1¢ treatment groups, so we instead apply the "midpoint" technique. For the weak demand treatments we report treatment effect estimates for both positive-positive and negative-negative demand treatment applications. Encouragingly, the estimates are all quite similar to one another, lying within 10 percent of the conventional treatment effect estimate.

2.4.5 Structural estimates

Under further assumptions, strong demand treatments permit structural estimation of demand-free model parameters (v), as well as ϕ and $E[h|h^L]$. Knowing v allows the researcher to make predictions about behaviour ab-

sent experimenter demand. Knowing ϕ allows them to quantify the importance of experimenter demand. Measuring beliefs can enable them to diagnose and eliminate the sources of latent demand effects. We illustrate how structural estimation can be performed using the real effort experiment. Because our model simply nests that of DellaVigna and Pope (2018) (DP), we follow their approach to structural estimation.[24]

DP estimate the following utility function (expressed in our notation):

$$v(a) = (s + \zeta)a - c(a) \tag{2.9}$$

The action a is effort, measured in points on the task, s is an intrinsic motivation parameter (workers may exert effort because they enjoy the task), and $c(a)$ is a cost of effort function. We assume the environment enters v only via the piece rate, so let $\zeta \in \{0, 1, 4\}$ be a real number. DP solve the first order condition and estimate the model parameters using nonlinear least squares (NLLS).[25]

Adding demand to this utility function gives:

$$U(a, \zeta) = (s + \zeta + \phi(\zeta)E[h|h^T, h^L(\zeta)])a - c(a) \tag{2.10}$$

with corresponding first-order condition

$$s + \zeta + \phi(\zeta)E[h|h^T, h^L(\zeta)] - c'(a^*(\zeta)) = 0 \tag{2.11}$$

DP consider two alternative forms for c: First, a power function $c(a) = ka^{1+\gamma}/(1+\gamma)$, yielding optimal effort equal to:

[24] We note that the structural analysis was not included in our pre-analysis plan.
[25] They also employ a minimum distance procedure. We stick to NLLS for brevity.

$$a^*(\zeta) = \left(\frac{s + \zeta + \phi(\zeta) E[h|h^T, h^L(\zeta)]}{k} \right)^{\frac{1}{\gamma}} \tag{2.12}$$

Second, an exponential form $c(a) = k \exp(\gamma a)/\gamma$, with effort level:

$$a^*(\zeta) = \frac{1}{\gamma} \log \left(\frac{s + \zeta + \phi(\zeta) E[h|h^T, h^L(\zeta)]}{k} \right) \tag{2.13}$$

We have seven treatment groups in total: neutral treatments with piece rates equal to 0 cents, 1 cent, and 4 cents per 100 points on the task; and positive and negative strong demand treatments in the 0 and 1 cent groups.[26] Noting that $E[h|h^L(\zeta)] = p^L(\zeta)h^L(\zeta) \in (-1, 1)$, we can treat it as a single parameter whose sign identifies h^L and whose magnitude identifies $p^L(\zeta)$. This leaves us with 10 parameters: s, k, γ, $\phi(0)$, $\phi(1)$, $\phi(4)$, $p^L(0)h^L(0)$, $p^L(1)h^L(1)$, $p^L(4)h^L(4)$, and p^T, so we need to impose some further restrictions.

First we assume that ϕ is fixed: $\phi(0) = \phi(1) = \phi(4) = \phi$, eliminating two parameters. In other words, varying incentives do not change the participants' desire to please the experimenter. Second, as in the previous section, we assume $p^T = 1$, which implies that $E[h|h^T, h^L] = h^T$. By assumption this is not justified for our weak demand treatments, so we focus on the strong treatments. We are left with seven parameters, s, k, γ, ϕ, $p^L(0)h^L(0)$, $p^L(1)h^L(1)$, and $p^L(4)h^L(4)$, and are therefore exactly identified. We additionally estimate a specification in which we restrict latent demand to depend only on whether monetary incentives are present, i.e. $p^L(1)h^L(1) = p^L(4)h^L(4)$.

While we use the same model as DP, identification comes from a different

[26] We also collected data using weak demand treatments, but we do not use it in this analysis a) because it was collected in a separate experiment and b) because for estimation we need to impose the parameter restriction $p^T = 1$, which we do not believe is satisfied in the weak treatments.

source. Under the assumption of no latent demand (as in DP), s, γ, and k are identified from the three *neutral treatment* groups. When latent demand is present, the model parameters (s, γ, k, ϕ) are identified from the *demand treatment* groups; with these in hand the neutral treatments allow us to back out the beliefs $p^L(\zeta)h^L(\zeta)$.

Full details of the estimation procedure, which follows DP, are provided in Appendix 2.B.9. We estimate equation (2.12) in logs and equation (2.13) in levels. Estimation results are presented in Table 2.6. Columns 1–3 correspond to the power cost function and columns 4–6 to the exponential cost function. In each case we first mirror DP by estimating s, γ, and k using only the neutral treatments, assuming that there is no latent demand.[27] Second, we include all treatment groups and impose that latent demand depends only on whether monetary incentives are present ($p^L(1)h^L(1) = p^L(4)h^L(4)$). Third, we allow latent demand to differ across all three incentive levels. Coefficients s and ϕ are measured in cents per 100 points. Therefore, $s = 1$ is interpreted as intrinsic motivation playing an equivalent role to an incentive of 1 cent per 100 points.

Our main finding is a nontrivial preference for pleasing the experimenter. Our estimates of ϕ take values in the range 0.2–0.3 and are similar across specifications. A value of 0.2 implies that moving from complete uncertainty ($E[h|h^L] = 0$) to complete certainty that high effort is desired ($E[h|h^L] = 1$) increases effort as much as increasing the incentive by 0.2 cents per 100 points.

Our estimates of $E[h|h^L]$ are mostly negative, consistent with latent demand decreasing effort. However, the estimates are noisy and typically not sig-

[27] There are some differences between our parameter estimates and DP's earlier work, which may reflect changes in the participant pool over time.

Table 2.6: Structural Estimates

	Power cost of effort			Exponential cost of effort		
	(1) Log Count	(2) Log Count	(3) Log Count	(4) Count	(5) Count	(6) Count
ϕ		0.175	0.249		0.205	0.300
		(0.092)	(0.095)		(0.079)	(0.066)
$h^L(0)p^L(0)$		-0.735	-0.516		-0.525	-0.187
		(0.172)	(0.303)		(0.191)	(0.249)
$h^L(>0)p^L(>0)$		-0.609			0.849	
		(2.194)			(1.799)	
$h^L(1)p^L(1)$			-0.473			0.155
			(1.110)			(0.694)
$h^L(4)p^L(4)$			-6.508			-6.600
			(3.360)			(1.963)
s	0.034	0.179	0.273	0.031	0.229	0.493
	(0.051)	(0.095)	(0.126)	(0.046)	(0.096)	(0.208)
k	4.7e-26	7.5e-24	6.5e-17	4.2e-08	2.1e-06	1.8e-04
	(3.1e-25)	(2.9e-23)	(3.1e-16)	(1.8e-07)	(3.7e-06)	(2.9e-04)
γ	7.260	6.583	4.433	6.5e-03	4.6e-03	2.3e-03
	(2.216)	(1.303)	(1.707)	(2.1e-03)	(8.7e-04)	(8.2e-04)
Observations	727	1691	1691	727	1691	1691
R-squared	0.122	0.166	0.166	0.167	0.204	0.206

Notes: This table uses data from the the real effort experiment on MTurk with strong demand treatments. Coefficients s and ϕ are measured in cents. s measures the respondents intrinsic motivation. ϕ measures the monetary value of acting according to the experimental objective. γ is the effort cost curvature and k is the scaling parameter. $h^L(\zeta)p^L(\zeta)$ is latent demand in incentive condition ζ. $h^L(>0)p^L(>0)$ is latent demand in the combined 1-cent and 4-cent incentive conditions. Robust standard errors in parentheses.

nificantly different from zero. We estimate that in the 4 cent treatment, $E[h|h^L(4)] \approx -6.5$, while the theory requires $E[h|h^L(4)] \in (-1, 1)$ (we note that the estimate is noisy and -1 lies well within the 95 percent confidence interval). This most likely reflects the fact that our demand treatments were only applied to the 0 and 1 cent treatment groups, so the effort cost function must be extrapolated far out of sample to estimate beliefs for the 4 cent group. We provide further discussion on this point, and an illustrative fig-

ure, in Appendix 2.B.9.

2.5 Properties of demand effects

In this section we examine some of the properties of demand effects and the assumptions underlying our approach. We begin with a discussion of Monotonicity, examining whether it holds first on average and then at the individual level. Second we turn to the central mechanism that drives behaviour in the model: changes in beliefs due signals about demand. Third, we consider the Bounding assumption. Although we cannot test it directly (since natural actions are not observed), we show that our bounds seem reasonable given existing evidence on responsiveness to a particular design feature – anonymity in the dictator game – that has been argued to potentially induce variation in demand. Fourth, we study heterogeneity in sensitivity to our demand treatments, focusing on four dimensions: incentives, gender, attention, and participant pool. These are cases where we might expect our Monotone Sensitivity assumption to hold, such that variation in sensitivity is informative about underlying variation in latent demand. Fifth, we examine the effect of our demand treatments on the variance and full distribution of actions.

2.5.1 Monotonicity

Monotonicity on average Our first theoretical assumption is Monotonicity: $a^+(\zeta) \geq a^L(\zeta) \geq a^-(\zeta)$. Panel C of Table 2.3 and Panel C of Table 2.4 examine this assumption for the subset of tasks in which we collected data

without applying demand treatments.[28] We estimate the following equation using the incentivised MTurk respondents, in which POS_i and NEG_i are dummy variables for the positive and negative demand treatments, and the no-demand condition is the reference group:

$$ZY_i = \pi_0 + \pi_1 POS_i + \pi_2 NEG_i + \varepsilon_i \tag{2.14}$$

We find strong support for Monotonicity in average actions. The strong demand treatments always moved average actions in the intended direction, and in most cases the differences are statistically significant. We find a significant negative response to negative weak demand in the investment game and a significant positive response to weak positive demand in the dictator game. Responses to the the positive demand treatment in the investment game and the negative demand treatment in the dictator game have the wrong signs but are close to zero and not statistically significant.

Testing monotonicity within-person Our seventh experiment uses a within-participant design, collecting data on behaviour first without, and then with a demand treatment. This allows us to examine Monotonicity directly at the individual level, and identify defiers, who try to do the opposite of the experimenter's wishes. Intuitively, by observing who increases and who decreases their action in response to a positive demand treatment, we can identify who is a complier and who is a defier. As discussed in section 2.2.4 "too much" defiance can invalidate our bounds.

[28] We have data for the dictator and investment games with weak and strong treatments, plus convex time budgets and real effort with only the strong treatments. Because the weak and strong treatments were applied in separate experiments, we analyze the data separately.

The design is as follows. MTurk participants were told that they would complete two tasks, and be paid according one of them, selected by chance. Half played the dictator game twice, and half the investment game twice. They first completed the task without any demand treatment, then again with the addition of a strong positive or negative demand treatment. We thus have four groups, split by dictator/investment game and positive/negative demand.

The model implies a simple interpretation of the data. Participants observe the first task, form a belief about h, and make a choice. They then observe the second task with the demand treatment, update their belief, and make a new choice. Strict compliers, with $\phi > 0$, will increase their action relative to task 1, strict defiers with $\phi < 0$ will decrease it, and those with $\phi = 0$ should take the same action in both tasks.[29]

Our main findings are captured by Figure 2.3, which plots actions from tasks 1 and 2. In the positive demand treatments, strict compliers lie above the 45 degree line, strict defiers lie below and those who did not change their action lie on the line. Only about 5 percent of our respondents are strict defiers. About 30 percent do not change their behaviour in response to our demand treatments, while the remaining 65 percent strictly comply with our demand treatments (proportions are similar across tasks). Thus we find very little evidence of defiance.

[29]The within design might fail to perfectly classify respondents, for two reasons. First, the theory assumes that ζ, and therefore the natural action, $a(\zeta)$, is independent of the demand treatment, h^T. This is a strong assumption in our within design, because it is clear that the response to h^T is part of the analysis, which could change participants' interpretation of ζ. However, if participants infer that our interest is in showing people respond to our demand treatments, compliers would increase and defiers to decrease their actions, in which case we would still arrive at the correct classification. Second, it might matter that participants have made a prior choice, either out of a concern for consistency (reducing responsiveness to our demand treatments) or a motive to conceal their defier/complier identity.

Table A.2 presents mean actions and sensitivities estimated from the within design and the equivalent objects from the earlier between-participants experiments. For the within experiment, "no demand" cells are computed from task 1, while demand treatment cells and sensitivities from task 2. The sensitivities are quantitatively very similar in the between and within designs. This is encouraging, as it suggests researchers can simply and cheaply obtain bounds using within-participant demand treatments, avoiding the need to recruit additional participants to apply our method.

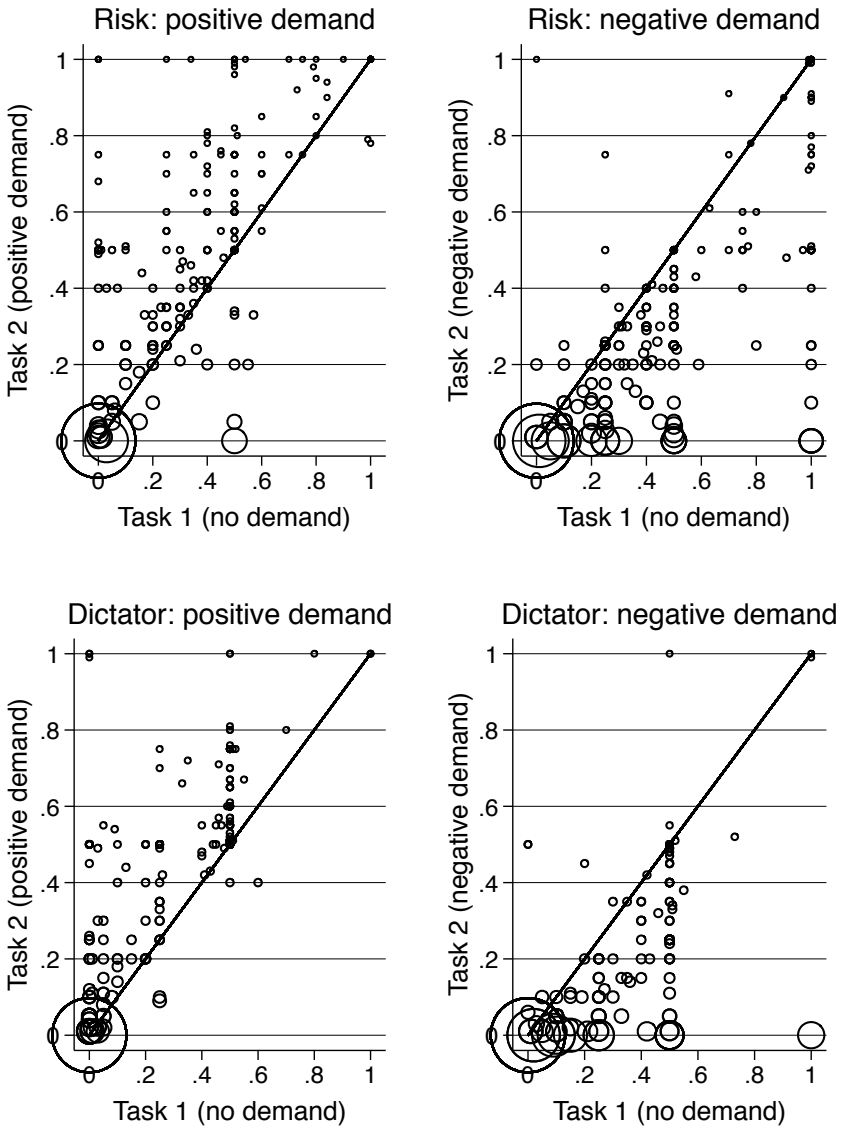
Within-participant data can be used to construct "defier-corrected" bounds.[30] These, with confidence intervals, are displayed in Table A.6. They are almost identical to the conventional bounds, reflecting the low rate of defiance and giving further comfort that defiance is quantitatively unimportant. Table A.5 reports raw actions separately for compliers and defiers.

2.5.2 Beliefs

The core mechanism in our model is that participants form beliefs about the experimenter's objective in response to implicit or explicit signals. We examine this assumption with simple, unincentivised belief data collected after participants had completed their experimental task. The purpose of the measures was a manipulation check, to ascertain that participants' beliefs responded as expected to the demand treatments. We asked two questions: "What do you think is the result that the researchers of this study want to find?," and "What do you think was the hypothesis of this research study?"

[30] For defiers, $a(\zeta) \in [a^+(\zeta), a^-(\zeta)]$ so, if the proportion of compliers is c the natural action lies in the interval $[cE[a^-(\zeta)|\phi \geq 0] + (1-c)E[a^+(\zeta)|\phi < 0], cE[a^+(\zeta)|\phi \geq 0] + (1-c)E[a^-(\zeta)|\phi < 0]]$. In practice one simply inverts the demand treatment variable for participants identified as defiers and computes bounds as before. The construction of defier-corrected bounds was not included in our pre-analysis plan.

Figure 2.3: Measuring defiance through a within-participant design



Notes: This figure uses MTurk data from experiment 7 and displays the scatterplot of responses in task 1 ("no demand" condition) and task 2 (demand condition). Points above the 45-degree line indicate an increase in the action, and points below the 45-degree line a decrease. The size of the rings is proportional to the number of observations.

Responses were binary: participants could respond that they thought the objective/hypothesis was either a high or low action.[31] We assume that

[31]One could collect richer belief measures and incentivize responses, but asking for fine-grained beliefs about *our own* motivations seemed quite unnatural, particularly as there was no objective truth against which to score. Our measures may of course be subject to their own demand bias.

participants report a high belief if their posterior ($E[h|h^L]$ or $E[h|h^T, h^L]$) is positive, and a low belief if negative, so the average response tells us the fraction of participants with high beliefs.

Results for incentivised MTurk respondents are presented in Tables A.8–A.9 in the Appendix. They confirm that our treatments moved average responses in the anticipated direction. Overall, the levels of beliefs and magnitudes of shifts in beliefs are similar for the strong and weak treatments, i.e. both were equally successful in fixing the sign of beliefs. In the theory, strong and weak treatments are equally effective at fixing the sign of beliefs if $p^T \geq p^L$, but stronger treatments lead to more extreme posteriors.[32]

2.5.3 Comparison of effect sizes

Is the bounding assumption reasonable? Although it is not directly testable, we compare our bounds to previous manipulations that have been hypothesised to induce demand effects. Our examples all come from the dictator game and include four studies that varied participants' degree of anonymity, and a study in Sierra Leone that varied the presence of a white foreigner.[33] We present effect sizes from these experiments and our own in Table A.11.

Sensitivity to our weak treatments (a 17 percent reduction in giving under negative versus positive demand) is very close to the average effect size

[32] $p^T \geq p^L$ also implies that *all* participants' beliefs should have the "correct" sign following a demand treatment. Not all of our participants reported correct beliefs following a demand treatment. This could be due to measurement error in our belief data, or, as we discuss in Appendix 2.B.3, participants might be inattentive to our demand treatments. If they are also inattentive to latent demand signals such participants do not threaten Bounding.

[33] Hoffman et al. (1994), Hoffman et al. (1996), Bolton et al. (1998) and Barmettler et al. (2012) study the effect of "double blind" versus "single blind" anonymity in dictator games, to our knowledge this is the complete set. Cilliers et al. (2015) study the effect of white foreigner presence.

across these 5 studies (around 21 percent reduction in giving in response to treatment), and our strong treatments comfortably bound this average (a 42 percent reduction). Considering individual studies, our weak bounds are close in magnitude to those from Bolton et al. (1998), Barmettler et al. (2012), and Cilliers et al. (2015), but smaller than those from Hoffman et al. (1994) and Hoffman et al. (1996). These two studies in particular, however, have been criticised for inducing potentially strong experimenter demand (Loewenstein, 1999), so may represent a scenario where the more conservative strong bounds are preferable. Their effect sizes are close to or a bit larger than (and not significantly different from) our strong bounds.

The exercise is of course only suggestive, since responses in these studies include direct effects of anonymity on behaviour as well as potential experimenter demand. Additionally, the studies we consider were conducted in the laboratory and differ in various other ways from our online setting. The results are nevertheless encouraging, in particular that our weak bounds seem to perform quite well.

2.5.4 Heterogeneity

Does sensitivity to demand treatments vary by design and participant characteristics? Here, we examine heterogeneous responses to our strong and weak demand treatments on four pre-specified dimensions: by whether choices are incentivised or hypothetical; gender; attentiveness; and participant pool (MTurk vs. representative online panel). Whether or not this heterogeneity can be interpreted as informative about differences in underlying latent demand depends upon whether Monotone Sensitivity holds for the environments under consideration, i.e. whether they belong to the

same comparison class. We show in Appendix 2.B.3 that variation in incentives, attention, and the preference for pleasing the experimenter, ϕ (which may differ by gender or participant pool), form valid bases for comparison classes.

incentivised vs. hypothetical choices In MTurk experiments 1 and 2 we randomly assigned participants to make either hypothetical or incentivised choices. In theory, we would expect higher sensitivity in hypothetical choice, as the cost of deviating from the natural action is lower. To test this prediction, we regress standardised actions on a dummy, POS_i, taking value one for the positive demand treatment and zero for the negative treatment; a dummy indicating incentivised choice, M_i; and their interaction:

$$ZY_i = \beta_0 + \beta_1 POS_i + \beta_2 M_i \times POS_i + \beta_3 M_i + \varepsilon_i \tag{2.15}$$

Results for the weak and strong demand treatments are presented in Table 2.7. Interestingly, participants making hypothetical or incentivised choices responded very similarly to experimenter demand, in each task and on average, and if anything sensitivity is slightly higher when incentivised.

Relatedly, we ask how sensitivity differs when we increase performance pay in the effort task. Reasonable assumptions would imply sensitivity is decreasing in performance pay (see Appendix 2.B.3). Table 2.4 shows that sensitivity to our strong treatments was around 3.5 times higher when effort was unincentivised, as predicted. We do not see the same pattern under the weak treatments, though this may simply reflect the fact that sensitivity to these treatments was low.

The mixed evidence on responsiveness to incentives is somewhat surpris-

ing. One possibility is that our incentivised choices still involve relatively low stakes, and that we would see a difference at higher stakes. Additionally, the theory allows ϕ to depend upon ζ and another possibility is that raising the stakes also raises participants' desire to please the experimenter (e.g. due to reciprocity). We see this as an interesting avenue for future work. Our results relate to previous work examining the effects of incentives on behaviour in the lab (Camerer et al., 1999).

Gender and attention We measure self-reported gender in all tasks on MTurk and in the representative panel, and attentiveness in all tasks except the effort task (since DP did not measure this variable). We define a participant as attentive if they passed an attention screener at the beginning of the task.[34] We estimate the following equation:

$$ZY_i = \beta_0 + \beta_1 POS_i + \beta_2 H_i + \beta_3 H_i \times POS_i + \varepsilon_i \tag{2.16}$$

where H_i is the dimension of heterogeneity of interest.

As can be seen in Table 2.7, we find that women respond more strongly to the strong demand treatments than men, with sensitivity around 0.15 standard deviations higher, but no significant difference for the weak treatments (where overall sensitivity and thus statistical power is lower). We interpret the evidence as suggestive of greater desire to please the experimenter among women, which relates to the literature on gender differences in preferences (Croson and Gneezy, 2009).

[34]We use the screener developed by Berinsky et al. (2014). It presents participants with a paragraph of text that appears to direct them to select their preferred online news sources from a list, but concealed in the text is an instruction to instead choose two specific options. The assumption is that attentive respondents read the question and follow the concealed instruction, while inattentive respondents do not.

Turning to attention, only 10 percent of MTurk respondents failed our screener, so we have little power to detect differences in sensitivity. Table 2.7 shows higher sensitivity (around 0.12 standard deviations) to our weak and strong manipulations among attentive participants, but these effects are not significant.

MTurk vs. representative online panel Some researchers are concerned that MTurk workers are experienced research participants and may behave differently than a more representative participant pool. In addition, MTurkers need to maintain a high work "acceptance" rating and may therefore be especially motivated to please the researcher (Berinsky et al., 2012). To address such concerns, and to test an additional dimension of heterogeneity, we replicated the MTurk dictator game and investment game experiments with respondents from a representative online panel, whose participants are less experienced in the types of tasks we consider. We used both weak and strong demand treatments, or no demand treatment. All choices were incentivised at the same stakes as in the MTurk experiments. Table 2.7 tests for differences in sensitivity between MTurk and representative survey participants, pooling tasks and for each task separately.

Representative panel participants responded very similarly to MTurk participants, with sensitivity on average 0.03 standard deviations higher (not significant) under both weak and strong treatments. There are some small differences in sensitivity to the strong treatments at the game level (significant at 10 percent for the dictator game), but little evidence of systematic differences between participant pools.

Table 2.7: Heterogeneity in response to weak and strong demand treatments (z-scored)

	All Games	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A: Weak - Design Characteristics												
Sensitivity × Incentive	0.073		0.149					-0.002				
	(0.085)		(0.125)					(0.115)				
Observations	1963		970					993				
Panel B: Weak - Respondent Characteristics												
Sensitivity × Male	0.038	-0.028	0.057	0.069	0.305	0.033	0.060	0.029	-0.089	0.003	0.257	-0.239
	(0.061)	(0.174)	(0.179)	(0.203)	(0.239)	(0.236)	(0.189)	(0.146)	(0.192)	(0.185)	(0.230)	(0.229)
Observations	4450	422	473	390	388	381	412	515	360	411	352	346
Sensitivity × Attention	0.119	-0.402	-0.077	0.434			0.226	0.585	0.094	0.368	0.116	-0.398
	(0.116)	(0.395)	(0.307)	(0.504)			(0.305)	(0.328)	(0.296)	(0.230)	(0.362)	(0.301)
Observations	3681	422	473	390			412	515	360	411	352	346
Sensitivity × Representative sample	0.032		0.032					0.032				
	(0.084)		(0.127)					(0.110)				
Observations	2125		1041					1084				
Panel C: Strong - Design Characteristics												
Sensitivity × Incentive	-0.007	-0.063	0.196					0.072				
	(0.080)	(0.132)	(0.116)					(0.121)				
Observations	2989	994	996					999				
Panel D: Strong - Respondent Characteristics												
Sensitivity × Male	-0.152	-0.212	-0.090	-0.382	0.075	0.005	-0.223	-0.201	-0.137	-0.144	0.098	-0.361
	(0.064)	(0.168)	(0.160)	(0.192)	(0.197)	(0.214)	(0.217)	(0.153)	(0.187)	(0.201)	(0.216)	(0.240)
Observations	4800	491	482	404	492	472	365	511	409	421	382	371
Sensitivity × Attention	0.117	0.319	0.471	-0.276			0.255	-0.024	-0.272	0.229	0.918	-0.091
	(0.140)	(0.393)	(0.401)	(0.414)			(0.358)	(0.530)	(0.394)	(0.538)	(0.409)	(0.311)
Observations	3836	491	482	404			365	511	409	421	382	371
Sensitivity × Representative sample	0.027		-0.121					0.176				
	(0.081)		(0.118)					(0.112)				
Observations	2184		1070					1114				

Notes: Outcome variables are z-scored at the task level. Panels A and C display heterogeneous treatment effects by design characteristics, i.e. whether choices are incentivised or hypothetical. Panels B and D display heterogeneous treatment effects by respondent characteristics: gender, attention and population. "Male" equals one for males, "attention" equals one if the respondent passed the attention screener, "representative sample" equals one for representative sample respondents.

2.5.5 Demand and the distribution of actions

We have focused on analysis of mean behavior, but other moments may respond to our demand treatments. For example, by aligning beliefs, they might reduce the variance of observed actions. Table A.12 shows that variance is very similar and in most cases slightly lower under the demand treatments relative to no demand. Figures A.2 and A.3 plot the cumulative distribution of actions for each task and demand treatment, showing that the demand treatments shift the full distribution of behavior. Encouragingly these shifts seem to almost always satisfy first-order stochastic dominance, consistent with Monotonicity.

2.6 Using the method in practice

We now provide some practical guidance on using the methods developed in this paper. First, we discuss settings in which demand treatments can be employed. Second, many of the applications in this paper have been to "levels" of behavior, so we list a few examples of other cases where one might be specifically interested in bounding levels. Third, we summarize the set of techniques and recommendations we have developed. Appendix 2.B.10 uses a diagram to work through an example of each technique.

We have two main settings in mind for applications. First, demand treatments can be applied in *experiments* in the laboratory, online, or in the field. We expect their primary use will be for the various robustness checks and estimation procedures we have outlined, but they can also be used for studying demand effects themselves. A natural next step in this agenda would be

to compare demand sensitivity in the lab and online, which may differ due to differences in attentiveness or social interaction with the experimenter. Second, they can readily be applied in *surveys*. Our estimates from hypothetical dictator games, convex time budgets and investment games, which are commonly used as survey questions, show that reasonable bounds are obtained even when choices are not incentivised. Applications include standalone surveys (e.g. on political views, inflation expectations, labour market outcomes) or field experiments, which often rely on survey data. For instance, participants might be told: "The researchers expect respondents who received the intervention (e.g. cash, bednets, education) to report more favourable outcomes".

While the majority of experiments are aimed at estimating treatment effects, researchers are often interested in mean responses in both surveys and experiments, and might be concerned about robustness. We provide a few examples. Policymakers might be intrinsically interested in levels of policy views about taxation or immigration; beliefs about these objects; willingness to contribute to public goods; inflation or growth expectations; consumption plans; or time use. In the lab, we are often interested in the level of giving in dictator games; offers and frequency of rejections in ultimatum games; competitiveness of specific subpopulations (e.g. men versus women); the amount of lying in coinflip games; or the degree of risk or ambiguity aversion (e.g. for calibrating models).

A further use of levels estimated in the lab or surveys is to predict behavior in other contexts (e.g. using risk, time or social preference measures to predict real-world behaviors). The extent to which these measures are predictive may be sensitive to demand effects, which can be thought of as a

form of measurement error. Our approach can be used to shed light on how important such errors might be. Within-subject applications even allow the researcher to measure and control for participant-level estimates of demand sensitivity.

We make the following recommendations on how to use demand treatments. First, in most studies we believe "weak" manipulations will give sufficiently conservative bounds, because explicit signals about the study hypothesis are likely to be more informative than implicit messages from the design in most cases. If potential demand confounds are a first-order concern, researchers may find stronger language, similar to our "strong" manipulations, helpful for further robustness. Our phrasings were chosen for broad applicability, but researchers with a specific application in mind may prefer to design their own demand treatments to best suit their setting.[35] With bounds in hand, researchers can compute demand-robust confidence intervals following Imbens and Manski (2004).

Second, demand treatments can be applied within-participant by adding a small number of questions or tasks to the end of a study. These are repetitions of questions or tasks presented earlier in the study, now including a demand treatment. Our estimates suggest that this approach yields similar bounds to a between-participant design, but is much less demanding of sample size. It also allows researchers to identify which participants are most sensitive to demand, and compute "defier-corrected" bounds. Third, we have shown how demand treatments can be used for point identification of treatment effects, applying same-signed demand treatments to the

[35]When bounding treatment effects, one could refer to the effect of interest in the demand treatment. For example, one could tell participants "You are in the high incentive treatment and will be compared with a group that has low incentives. We expect that incentives will increase effort".

treatment and control group. If demand treatments are "sufficiently informative" this approach can eliminate biases due to differences in beliefs, and any remaining bias due to differences in behavioural responsiveness can be tested for. We have also shown how sufficiently informative demand treatments can be used for structural identification of models, by plausibly eliminating nuisance parameters due to unobservable beliefs.

Fourth, in a study with many treatment arms, adding all of the possible demand manipulations may become impractical. In such settings, researchers could add demand manipulations to a subset of groups, and then compare treatment effect magnitudes to demand sensitivity measured in those groups. When an experiment features many different and complicated choices, researchers may find it worthwhile to consider what overarching beliefs could affect their estimates (for example, participants might believe that they should misreport their valuations in willingness-to-pay elicitation), and target those with demand treatments, rather than manipulating individual actions. Finally, researchers conducting similar experiments to those in this paper may find our estimates useful for benchmarking purposes.

2.7 Conclusion

We propose a technique for assessing the robustness of experimental results to demand effects. We deliberately induce demand in a structured way to measure its influence and to construct bounds on demand-free behavior and treatment effects. We formalize the intuition behind the procedure with a simple model in which participants form beliefs about the experimental objective and gain utility from conforming to it. Bounds are obtained by intentionally manipulating those beliefs.

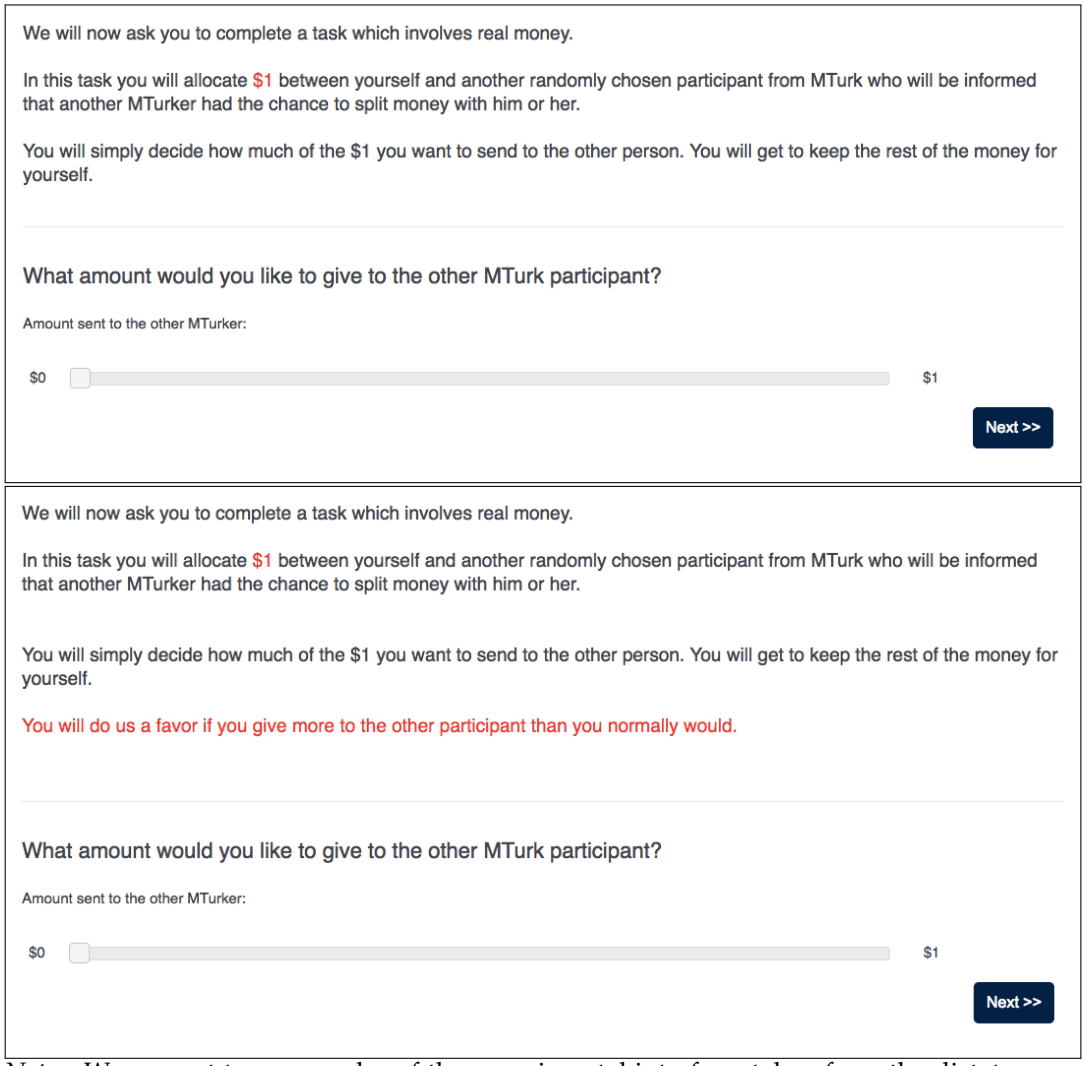
Across eleven canonical experimental tasks we find modest responses to demand manipulations that explicitly signal the researcher's hypothesis, with bounds averaging around 0.13 standard deviations in width. We argue that these treatments reasonably bound the magnitude of demand effects in typical experiments, so our findings give cause for optimism. Using stronger manipulations we show how to obtain demand-robust point estimates of treatment effects, and analyze demand effects structurally. In a real effort task with incentives of 1 cent per 100 "points," we estimate a utility of pleasing the experimenter of around 0.2 cents per 100 points.

Future work might employ similar treatments to study how to mitigate demand in experiments, for example by examining how demand sensitivity varies with features of the environment. One avenue for further exploration is the effect of incentives, given the central role they play in experiments.

Appendices of Chapter 2

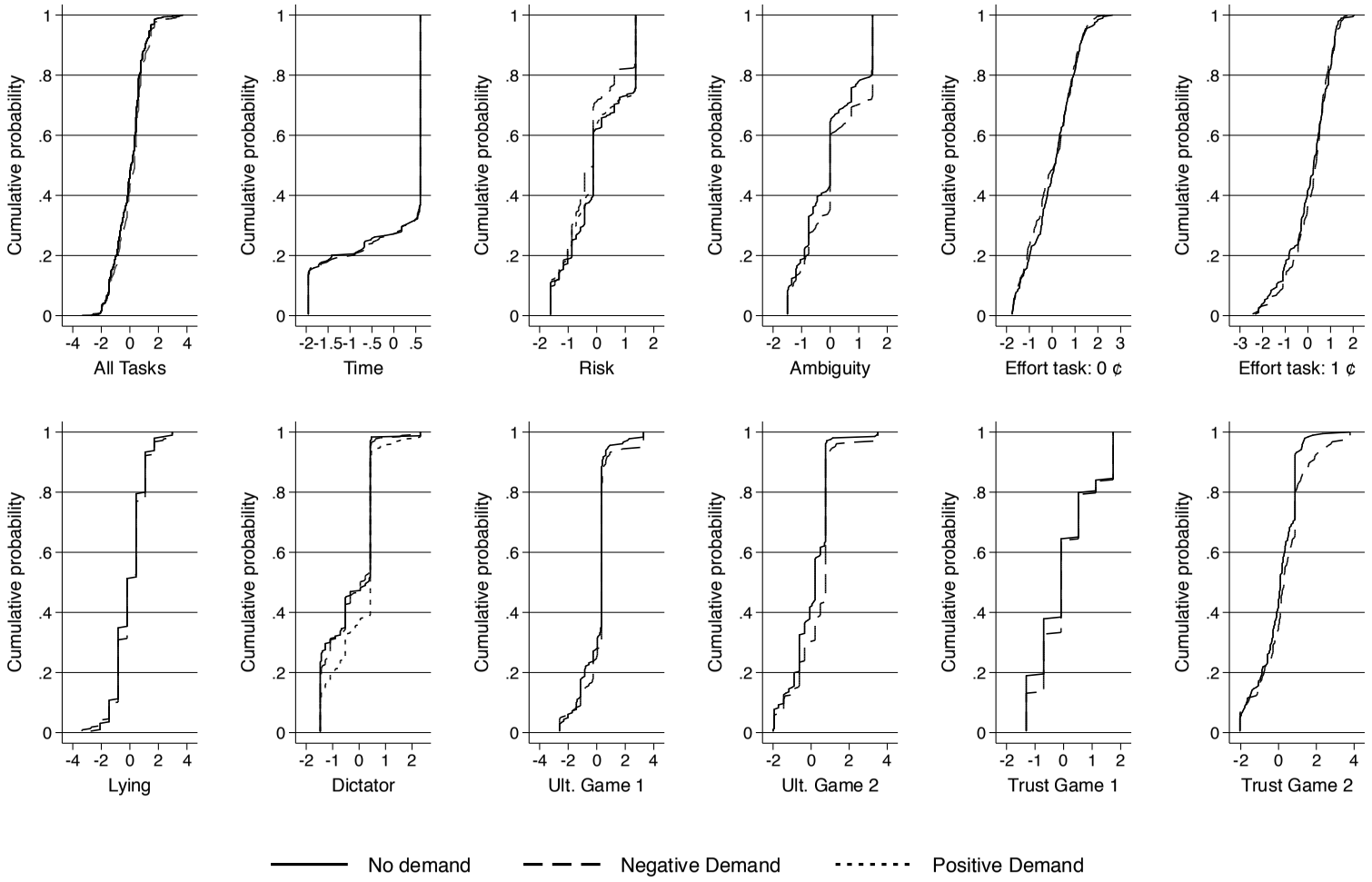
2.A Additional Figures and Tables

Figure A.1: Interface example



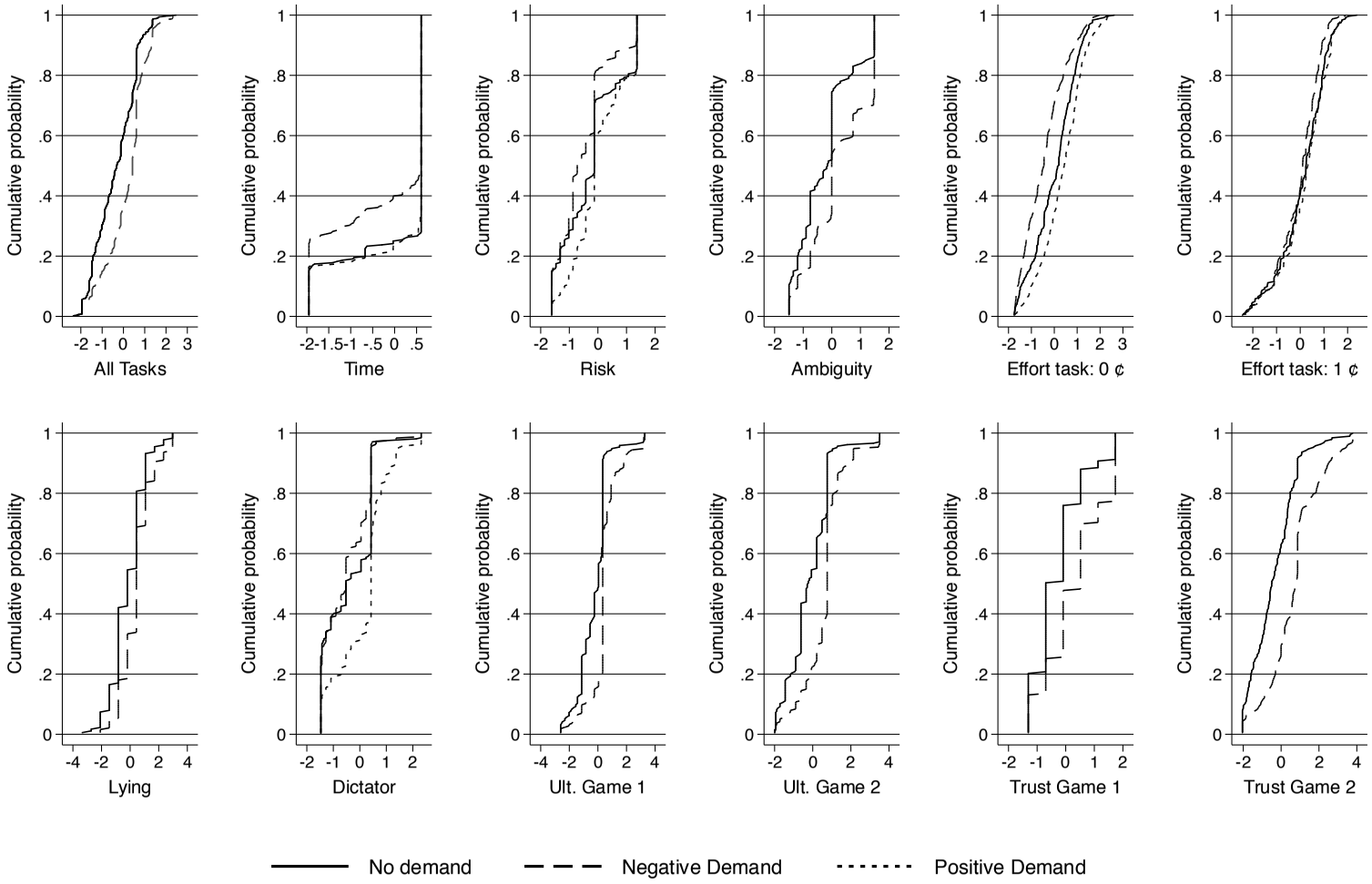
Notes: We present two examples of the experimental interface, taken from the dictator game. The first frame corresponds to the real stakes, "no demand" condition, and the second frame to the real stakes, positive demand condition.

Figure A.2: Distribution of z-scored actions by task and demand treatment, weak treatments



Notes: This figure uses data from incentivised MTurk respondents with weak demand treatments, and displays the cumulative distribution function of z-scored actions by task and demand treatment arm.

Figure A.3: Distribution of z-scored actions by task and demand treatment, strong treatments



Notes: This figure uses data from incentivised MTurk respondents with strong demand treatments, and displays the cumulative distribution function of z-scored actions by task and demand treatment arm.

Table A.1: Controlling for demand

	Conventional	Sensitivity difference	Strong	Weak	
	Treatment Effect	(Strong 1-cent - Strong 0-cent)	Midpoint	positive-positive	negative-negative
Count	540.720	430.009	494.774	588.270	530.001
	(66.763)	(89.477)	(49.321)	(61.499)	(64.532)
Count (z-scored)	0.686	0.546	0.628	0.747	0.673
	(0.085)	(0.114)	(0.063)	(0.078)	(0.082)

Notes: This table uses data from the real effort experiments (experiment 3 and experiment 6). We follow the "controlling for demand" procedure outlined in Section 2.4.4 to estimate the treatment effect of incentives on effort provision. Column (1) shows the conventional treatment effect estimate (data from experiment 3). Column (2) tests for differences in sensitivity to our strong demand treatments between the 0-cent and 1-cent groups, and finds a significant difference. Therefore in column (3) we apply the "midpoint" technique with strong demand treatments to estimate the treatment effect. Columns (4) and (5) approximate the treatment effect using same-signed weak demand treatments. We apply the "ironing" procedure described in section 2.4.2 when constructing these estimates. Count is the raw-score of points scored in the real effort task. Count (z-scored) uses the mean and standard deviation from the negative demand condition. Robust standard errors in parentheses. Note that strong and weak treatment data were collected in separate experiments.

Table A.2: Results from the Within Design

	Dictator			Risk		
	Within	Between	Difference	Within	Between	Difference
Panel A: Unconditional Means						
Positive demand	0.384	0.434	-0.050	0.560	0.550	0.010
	(0.017)	(0.015)	(0.023)	(0.021)	(0.020)	(0.029)
No demand	0.273	0.282	-0.010	0.448	0.466	-0.018
	(0.011)	(0.015)	(0.019)	(0.015)	(0.022)	(0.027)
Negative demand	0.195	0.251	-0.056	0.318	0.373	-0.055
	(0.014)	(0.014)	(0.020)	(0.019)	(0.019)	(0.027)
Panel B: Sensitivity (positive - negative)						
Raw data	0.189	0.183	0.006	0.242	0.177	0.065
	(0.022)	(0.021)	(0.031)	(0.029)	(0.027)	(0.040)
Z-score	0.794	0.745	0.048	0.709	0.520	0.188
	(0.093)	(0.086)	(0.127)	(0.084)	(0.080)	(0.116)
Panel C: Monotonicity						
Positive - Neutral (z-score)	0.514	0.617	-0.103	0.377	0.248	0.129
	(0.044)	(0.088)	(0.129)	(0.041)	(0.087)	(0.124)
Negative - Neutral (z-score)	-0.380	-0.128	-0.251	-0.427	-0.272	-0.155
	(0.045)	(0.086)	(0.123)	(0.042)	(0.084)	(0.119)
Observations	499	770	1269	500	728	1228

Notes: This table uses data from the within design (experiment 7) and incentivised choices from the dictator game and the investment game in experiment 1. These experiments employ strong demand treatments. Panel A displays the unconditional means by task and demand treatment arm. Panel B displays the estimates of sensitivity. Panel C tests Monotonicity. Note that estimates from Panel C do not add up to the sensitivity estimates from Panel B as sensitivity is estimated between participants while monotonicity tests are within-participant.

Table A.3: Confidence intervals for bounds on natural actions

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A: Weak Demand											
Interval	[0.766, 0.770]	[0.472, 0.524]	[0.499, 0.557]	[0.343, 0.331]	[0.469, 0.484]	[0.469, 0.484]	[0.318, 0.382]	[0.443, 0.470]	[0.362, 0.413]	[0.430, 0.455]	[0.348, 0.398]
95% CI on interval	[0.716, 0.821]	[0.438, 0.561]	[0.459, 0.597]	[0.316, 0.358]	[0.448, 0.504]	[0.510, 0.557]	[0.296, 0.405]	[0.422, 0.493]	[0.342, 0.436]	[0.387, 0.494]	[0.328, 0.426]
95% CI on parameter	[0.724, 0.812]	[0.445, 0.553]	[0.468, 0.588]	[0.320, 0.353]	[0.452, 0.499]	[0.514, 0.553]	[0.301, 0.400]	[0.427, 0.488]	[0.346, 0.431]	[0.396, 0.486]	[0.332, 0.420]
Observations	422	739	390	388	381	412	758	360	411	352	346
Panel B: Strong Demand											
Interval	[0.659, 0.795]	[0.373, 0.550]	[0.428, 0.583]	[0.255, 0.405]	[0.449, 0.492]	[0.449, 0.492]	[0.251, 0.434]	[0.404, 0.520]	[0.337, 0.474]	[0.350, 0.535]	[0.288, 0.469]
95% CI on interval	[0.612, 0.834]	[0.342, 0.583]	[0.391, 0.622]	[0.236, 0.424]	[0.432, 0.511]	[0.487, 0.626]	[0.227, 0.459]	[0.381, 0.541]	[0.314, 0.496]	[0.314, 0.574]	[0.263, 0.498]
95% CI on parameter	[0.622, 0.826]	[0.349, 0.576]	[0.399, 0.613]	[0.240, 0.420]	[0.436, 0.507]	[0.493, 0.622]	[0.232, 0.454]	[0.386, 0.536]	[0.319, 0.491]	[0.322, 0.565]	[0.269, 0.491]
Observations	727	728	404	731	714	365	770	409	421	382	371

Notes: This table uses data from incentivised MTurk respondents with strong and weak demand treatments. It first presents estimated bounds on the natural action, then 95 percent confidence intervals on those bounds, then 95 percent confidence intervals on the parameter (natural action) contained in the bounds.

Table A.4: Confidence intervals for bounds on treatment effects

	Treatment Effect: Score in Effort Task
Weak treatments	
Interval	[530.001, 588.270]
95% CI on interval	[410.310, 701.645]
95% CI on parameter	[434.271, 678.736]
Observations	769
Strong treatments	
Interval	[177.421, 948.978]
95% CI on interval	[74.817, 1054.492]
95% CI on parameter	[97.479, 1031.187]
Observations	1445

Notes: This table uses data from incentivised MTurk respondents with weak and strong demand treatments (experiments 3 and 6). It first presents estimated bounds on the treatment effect of incentives on effort, then 95 percent confidence intervals on those bounds, then 95 percent confidence intervals on the parameter (treatment effect) contained in the bounds.

Table A.5: Results from the Within Design: Compliers and Defiers

	Dictator			Risk		
	All	Compliers	Defiers	All	Compliers	Defiers
Positive - Neutral (z-score)	0.514	0.777	-0.402	0.377	0.704	-0.601
	(0.044)	(0.055)	(0.122)	(0.041)	(0.052)	(0.100)
Observations	265	179	7	247	146	16
Negative - Neutral (z-score)	-0.380	-0.796	1.028	-0.427	-0.721	0.529
	(0.045)	(0.059)	(0.329)	(0.042)	(0.049)	(0.199)
Observations	234	122	8	253	161	16

Notes: This table uses data from the within design (experiment 7). The outcome variable is the change in standardised action between task 1 and task 2. We separately present the results for the whole sample, compliers, and defiers.

Table A.6: Within Design: Defier-corrected bounds and confidence intervals

	Risk	Dictator
Panel A: Standard Bounds		
Interval	[0.318, 0.560]	[0.195, 0.384]
95% CI on interval	[0.286, 0.595]	[0.172, 0.412]
95% CI on parameter	[0.293, 0.587]	[0.177, 0.406]
Observations	500	499
Panel B: Adjusted Bounds		
Interval	[0.308, 0.571]	[0.185, 0.392]
95% CI on interval	[0.277, 0.606]	[0.163, 0.420]
95% CI on parameter	[0.284, 0.598]	[0.168, 0.414]
Observations	500	499

Notes: This table uses data from the within design (experiment 7). In Panel A we compute our standard bounds and confidence intervals. In Panel B we compute the adjusted bounds which take into account defier behavior.

Table A.7: Belief about the experimental objective in response to the weak demand treatments

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.832	0.757	0.760	0.788	0.979	0.779	0.540	0.698	0.688	0.612	0.669
	(0.026)	(0.028)	(0.031)	(0.030)	(0.010)	(0.028)	(0.032)	(0.034)	(0.032)	(0.036)	(0.038)
No demand		0.620					0.321				
		(0.030)					(0.030)				
Negative demand	0.603	0.370	0.330	0.277	0.358	0.467	0.234	0.238	0.383	0.112	0.083
	(0.034)	(0.031)	(0.034)	(0.032)	(0.035)	(0.036)	(0.026)	(0.032)	(0.034)	(0.024)	(0.020)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.229	0.388	0.430	0.511	0.621	0.312	0.306	0.461	0.304	0.500	0.585
	(0.042)	(0.042)	(0.046)	(0.044)	(0.036)	(0.046)	(0.041)	(0.047)	(0.047)	(0.044)	(0.043)
Z-score	0.471	0.776	0.909	1.117	1.240	0.627	0.678	0.994	0.633	1.092	1.417
	(0.087)	(0.084)	(0.096)	(0.095)	(0.073)	(0.092)	(0.091)	(0.101)	(0.098)	(0.095)	(0.104)
		[0.001]					[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)		0.274					0.485				
		(0.082)					(0.096)				
		[0.001]					[0.001]				
Negative - Neutral (z-score)		-0.501					-0.193				
		(0.087)					(0.088)				
		[0.001]					[0.009]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Notes: This table uses data from incentivised MTurk respondents with weak demand treatments. The outcome variables take value one if the respondents believed that the experimenter wanted a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task.

Table A.8: Belief about the experimental objective in response to the strong demand treatments

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.802	0.705	0.701	0.776	0.942	0.815	0.651	0.572	0.664	0.407	0.385
	(0.025)	(0.030)	(0.032)	(0.027)	(0.015)	(0.028)	(0.029)	(0.034)	(0.032)	(0.035)	(0.036)
No demand	0.720	0.537		0.485	0.888		0.355				
	(0.029)	(0.032)		(0.032)	(0.020)		(0.030)				
Negative demand	0.622	0.424	0.335	0.296	0.511	0.562	0.244	0.309	0.357	0.295	0.217
	(0.031)	(0.031)	(0.033)	(0.029)	(0.033)	(0.037)	(0.028)	(0.033)	(0.033)	(0.034)	(0.030)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.181	0.281	0.366	0.480	0.431	0.252	0.407	0.263	0.306	0.112	0.168
	(0.040)	(0.043)	(0.046)	(0.039)	(0.036)	(0.047)	(0.040)	(0.047)	(0.047)	(0.049)	(0.047)
Z-score	0.372	0.563	0.773	1.050	0.861	0.507	0.901	0.567	0.637	0.245	0.406
	(0.083)	(0.087)	(0.098)	(0.086)	(0.073)	(0.094)	(0.089)	(0.102)	(0.097)	(0.106)	(0.114)
	[0.001]	[0.001]		[0.001]	[0.001]		[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)	0.169	0.337		0.635	0.108		0.654				
	(0.079)	(0.088)		(0.092)	(0.050)		(0.092)				
	[0.023]	[0.001]		[0.001]	[0.011]		[0.001]				
Negative - Neutral (z-score)	-0.203	-0.226		-0.415	-0.754		-0.247				
	(0.089)	(0.089)		(0.095)	(0.077)		(0.090)				
	[0.022]	[0.003]		[0.001]	[0.001]		[0.002]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Notes: This table uses data from incentivised MTurk respondents with strong demand treatments. The outcome variables take value one if the respondents believed that the experimenter wanted a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task.

Table A.9: Belief about the experimental hypothesis in response to the weak demand treatments

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.788	0.749	0.704	0.782	0.963	0.871	0.464	0.726	0.732	0.628	0.682
	(0.028)	(0.028)	(0.033)	(0.030)	(0.014)	(0.023)	(0.032)	(0.033)	(0.031)	(0.036)	(0.038)
No demand		0.534					0.160				
		(0.031)					(0.024)				
Negative demand	0.458	0.261	0.222	0.231	0.326	0.528	0.106	0.354	0.359	0.296	0.182
	(0.034)	(0.029)	(0.030)	(0.030)	(0.034)	(0.036)	(0.019)	(0.036)	(0.034)	(0.035)	(0.028)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.331	0.488	0.482	0.552	0.637	0.343	0.358	0.373	0.372	0.333	0.500
	(0.044)	(0.040)	(0.044)	(0.042)	(0.037)	(0.042)	(0.037)	(0.049)	(0.046)	(0.050)	(0.047)
Z-score	0.681	0.978	0.982	1.244	1.286	0.706	0.836	0.825	0.750	0.731	1.161
	(0.092)	(0.080)	(0.090)	(0.096)	(0.074)	(0.087)	(0.086)	(0.108)	(0.092)	(0.110)	(0.109)
		[0.001]					[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)		0.431					0.708				
		(0.084)					(0.092)				
		[0.001]					[0.001]				
Negative - Neutral (z-score)		-0.547					-0.128				
		(0.084)					(0.071)				
		[0.001]					[0.024]				
Observations	422	739	390	388	381	412	758	360	411	352	346

Notes: This table uses data from incentivised MTurk respondents with weak demand treatments. The outcome variables take value one if the respondents believed that the experimenter expected a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task.

Table A.10: Belief about the experimental hypothesis in response to the strong demand treatments

	Belief: Time	Belief: Risk	Belief: Ambiguity Aversion	Belief: Effort 0 cent bonus	Belief: Effort 1 cent bonus	Belief: Lying	Belief: Dictator Game	Belief: Ult. Game 1	Belief: Ult. Game 2	Belief: Trust Game 1	Belief: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.727	0.595	0.593	0.727	0.934	0.788	0.454	0.670	0.659	0.578	0.588
	(0.028)	(0.033)	(0.034)	(0.029)	(0.016)	(0.030)	(0.030)	(0.032)	(0.032)	(0.035)	(0.037)
No demand	0.682	0.484		0.423	0.855		0.143				
	(0.030)	(0.032)		(0.032)	(0.023)		(0.022)				
Negative demand	0.639	0.420	0.400	0.267	0.576	0.625	0.186	0.284	0.435	0.290	0.243
	(0.031)	(0.031)	(0.035)	(0.028)	(0.033)	(0.037)	(0.025)	(0.032)	(0.035)	(0.034)	(0.031)
Panel B: Sensitivity (Positive - Negative)											
Raw data	0.089	0.175	0.193	0.459	0.358	0.163	0.268	0.386	0.224	0.288	0.345
	(0.042)	(0.045)	(0.049)	(0.040)	(0.036)	(0.047)	(0.039)	(0.046)	(0.047)	(0.049)	(0.048)
Z-score	0.183	0.351	0.393	1.036	0.722	0.336	0.624	0.855	0.451	0.634	0.801
	(0.087)	(0.090)	(0.100)	(0.091)	(0.074)	(0.097)	(0.092)	(0.101)	(0.095)	(0.107)	(0.112)
	[0.118]	[0.001]		[0.001]	[0.001]		[0.001]				
Panel C: Monotonicity											
Positive - Neutral (z-score)	0.093	0.222		0.685	0.159		0.725				
	(0.085)	(0.091)		(0.097)	(0.056)		(0.087)				
	[0.268]	[0.015]		[0.001]	[0.001]		[0.001]				
Negative - Neutral (z-score)	-0.090	-0.128		-0.350	-0.563		0.101				
	(0.090)	(0.089)		(0.096)	(0.080)		(0.077)				
	[0.268]	[0.052]		[0.001]	[0.001]		[0.069]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Notes: This table uses data from incentivised MTurk respondents with strong demand treatments. The outcome variables take value one if the respondents believed that the experimenter expected a high action. Panel A displays mean beliefs with standard errors in the positive, negative and no-demand conditions respectively. Panel B presents the raw and z-scored sensitivity of beliefs to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when "no demand" choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets, adjusting across tests within each task.

Table A.11: Overview of studies varying anonymity in dictator games

Study	Description of the treatment	Sample	Sample Size	Stake Size	Effect Size	Statistical significance
Hoffman et al. (1994)	Double blind compared to single blind	Student sample from the University of Arizona	101	$10	61 percent reduction in giving	p<0.01
Hoffman et al. (1996)	Double blind compared to single blind	Student sample from the University of Arizona	114	$10	37 percent reduction in giving	p<0.01
Bolton et al. (1998)	Double blind compared to single blind	Student sample at Penn State University	60	$5	22 percent increase in giving	p>0.1
Barmettler et al. (2012)	Double blind compared to single blind	Student samples from the University of Zurich (UZH)	103	20 Swiss Frank ($22)	16.8 percent reduction in giving	p>0.1
Cilliers et al. (2015)	Presence of non-foreign experimenter vs.presence of a white foreign experimenter	Poor households from Sierra Leone	708	4000 Leones (approximately $1)	16 percent reduction in giving	p<0.01
Our estimates based on demand treatments						
de Quidt et al. (2018)	Weak negative demand treatment compared to weak positive demand treatment	MTurk respondents	515	$1	17 percent reduction in giving	p<0.01
de Quidt et al. (2018)	Strong negative demand treatment compared to strong positive demand treatment	MTurk respondents	511	$1	42 percent reduction in giving	p<0.01

Notes: This table provides an overview of dictator game studies which vary the anonymity of experimenter-subject interactions and the presence of a foreign (white) experimenter. Our estimates of treatment effects for the studies by Hoffman et al. (1994) and Hoffman et al. (1996) are based on inspection of the cumulative distribution functions and and probability distribution functions reported in the paper (details of our calculations are available upon request). These papers did not report mean behaviour across treatment arms. In Hoffman et al. (1994) we compare behaviour in "Double Blind treatment 1" and "Double Blind treatment 2" to behaviour in the "Dictator random entitlement, exchange". In Hoffman et al. (1996) we compare behaviour in "Double Blind treatment 1" and "Double Blind treatment 2" to behaviour in the "Single Blind 1" condition. In Bolton et al. (1998) we compare behaviour in the "Anonymity" condition to behaviour in the "6card1game" condition. In Barmettler et al. (2012) we compare behaviour in the "Double Anonymity" condition to behaviour in the "Single Anonymity" condition. In Cilliers et al. (2015) we compare behaviour when a white foreigner was or was not present in the session. The average reduction in giving across the studies using equal weights is a 21.76 percent, or 20.37 percent when weighted by sample size.

Table A.12: Overview of standard deviations across tasks

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
Panel A: Weak Demand											
Positive demand	0.385	0.348	0.341	0.193	0.165	0.170	0.222	0.189	0.194	0.314	0.217
No demand	.	0.339	.	.	.	.	0.234	.	.	.	.
Negative demand	0.389	0.317	0.334	0.190	0.178	0.158	0.226	0.170	0.182	0.329	0.172
Panel B: Strong Demand											
Positive demand	0.379	0.331	0.340	0.177	0.179	0.172	0.267	0.184	0.202	0.334	0.234
No demand	0.386	0.340	.	0.182	0.184	.	0.246	.	.	.	.
Negative demand	0.437	0.322	0.319	0.176	0.162	0.183	0.229	0.189	0.209	0.291	0.205

Notes: This table uses data from incentivised MTurk respondents with weak and strong demand treatments and displays the standard deviations across the different demand treatment arms.

2.B Theoretical Appendix

2.B.1 Proof of Proposition 1 (Monotonicity)

We require that $a^+(\zeta) \geq a^L(\zeta) \geq a^-(\zeta)$. We are therefore interested in the sign of $\phi(E[h|h^T, h^L(\zeta)] - E[h|h^L(\zeta)])$. We have:

$$\begin{aligned}\phi(E[h|h^T, h^L(\zeta)] - E[h|h^L(\zeta)]) &= \phi\left(\frac{h^L(\zeta)p^L(\zeta) + h^T p^T}{1 + h^L(\zeta)p^L(\zeta)h^T p^T} - h^L(\zeta)p^L(\zeta)\right) \\ &= \phi h^T p^T \frac{(1 - h^L(\zeta)^2 p^L(\zeta)^2)}{1 + h^L(\zeta)p^L(\zeta)h^T p^T}\end{aligned}$$

Because we assumed that $p^L(\zeta) < 1$, this expression has the same sign as $\phi h^T p^T$. We want to show that $\phi(E[h|h^T = 1, h^L(\zeta)] - E[h|h^L(\zeta)]) \geq 0$ and $\phi(E[h|h^T = -1, h^L(\zeta)] - E[h|h^L(\zeta)]) \leq 0$. This follows trivially when $p^T = 0$. When $p^T > 0$ if follows if and only if $\phi \geq 0$.

2.B.2 Proof of Proposition 2 (Bounding)

In the Bayesian model, given $\phi \geq 0$ (Monotonicity), the action is larger or smaller than $a(\zeta)$ when $\phi E[h|h^T, h^L] \geq 0$ or $\phi E[h|h^T, h^L] \leq 0$ respectively. Given that $\phi \geq 0$, we need $E[h|h^T = 1, h^L] \geq 0$ and $E[h|h^T = -1, h^L] \leq 0$. This is guaranteed if h^T and h^L have the same sign, so we simply need to check whether it holds when the demand treatment and latent demand are in opposite directions, i.e. $E[h|h^T = 1, h^L = -1] \geq 0$ and $E[h|h^T = -1, h^L = 1] \leq 0$. Given our restriction $p^L(\zeta) < 1$, inspection of (2.7) reveals that these conditions hold if and only if $p^T \geq p^L(\zeta)$, i.e. the decision-maker

perceives the demand treatment as at least as informative about h as the latent demand signal.

2.B.3 Conditions for Monotone Sensitivity

Assumption 3 (Monotone Sensitivity) assumes that sensitivity $S(\zeta) = a^+(\zeta) - a^-(\zeta)$ is (strictly) monotone in the size of the latent demand effect $\left|a^L(\zeta) - a(\zeta)\right|$. Here we examine cases under which that is and is not the case. We assume throughout that Assumptions 1 and 2 hold.

Variation driven by ϕ. We are interested in how ϕ affects latent demand $(d\left|a^L(\zeta) - a(\zeta)\right|/d\phi)$ and sensitivity $(dS(\zeta)/d\phi)$. From (2.5) we obtain:

$$\frac{d(a^L(\zeta) - a(\zeta))}{d\phi} = -\frac{h^L(\zeta)p^L(\zeta)}{v_{11}(a^L(\zeta), \zeta)}$$

which has the same sign as $h^L(\zeta)$, allowing us to write $\frac{d\left|a^L(\zeta)-a(\zeta)\right|}{d\phi} = -\frac{p^L(\zeta)}{v_{11}(a^L(\zeta),\zeta)} \geq 0$.

Turning to sensitivity, we have:

$$\begin{aligned}\frac{dS(\zeta)}{d\phi} &= \frac{da^+(\zeta)}{d\phi} - \frac{da^-(\zeta)}{d\phi} \\ &= -\frac{1}{v_{11}(a^+(\zeta), \zeta)}\frac{h^L(\zeta)p^L(\zeta) + p^T}{1 + h^L(\zeta)p^L(\zeta)p^T} + \frac{1}{v_{11}(a^-(\zeta), \zeta)}\frac{h^L(\zeta)p^L(\zeta) - p^T}{1 - h^L(\zeta)p^L(\zeta)p^T}\end{aligned}$$

By Assumption 2, $h^L(\zeta)p^L(\zeta) + p^T \geq 0$ and $h^L(\zeta)p^L(\zeta) + p^T \leq 0$, so both terms are positive, i.e. $\frac{dS(\zeta)}{d\phi} \geq 0$. Therefore Monotone Sensitivity holds and any set of environments that differ only in ϕ constitutes a comparison class, i.e. for such environments, sensitivity is informative about the magnitude

of latent demand effects.

Example 1. *Suppose participant pool A is more concerned for pleasing the experimenter than participant pool B. Then latent demand effects and sensitivity will be larger in magnitude in participant pool A.*

Variation driven by v. Suppose that ζ can be separated into a parameter, z, and a remainder term, ζ', that v is differentiable in z and that ϕ, h^L and p^L do not depend on z. z could be a preference parameter (e.g. risk aversion) or a design parameter (e.g. the scale of incentives). We write $U(a, \zeta', z) = v(a, \zeta', z) + a\phi(\zeta')E[h|\zeta']$ and modify the first-order conditions accordingly.

$$\begin{aligned}
\frac{d(a^L(\zeta', z) - a(\zeta', z))}{dz} &= \frac{da^L(\zeta', z)}{dz} - \frac{da(\zeta', z)}{dz} \\
&= -\left[\frac{v_{13}(a^L(\zeta', z), \zeta', z)}{v_{11}(a^L(\zeta', z), \zeta', z)} - \frac{v_{13}(a(\zeta', z), \zeta', z)}{v_{11}(a(\zeta', z), \zeta', z)}\right] \\
\frac{dS(\zeta', z)}{dz} &= -\left[\frac{v_{13}(a^+(\zeta', z), \zeta', z)}{v_{11}(a^+(\zeta', z), \zeta', z)} - \frac{v_{13}(a^-(\zeta', z), \zeta', z)}{v_{11}(a^-(\zeta', z), \zeta', z)}\right]
\end{aligned}$$

It is clear from inspecting these conditions that we need to know how v_{13}/v_{11} varies with a, i.e.:

$$\frac{d\frac{v_{13}(a,\zeta',z)}{v_{11}(a,\zeta',z)}}{da} = \frac{v_{11}(a, \zeta', z)v_{113}(a, \zeta', z) - v_{111}(a, \zeta', z)v_{13}(a, \zeta', z)}{v_{11}(a, \zeta', tz)}$$

It is difficult to make general statements about these objects for general utility functions, so we focus attention on two special cases of interest.

Multiplicative separability. Suppose that $v(a, \zeta', z) = \nu(a, \zeta')f(z)$ and define z such that $f'(z) > 0$. Then

$$\begin{aligned}\frac{d\left(a^L(\zeta', z) - a(\zeta', z)\right)}{dz} &= -f'(z)\left[\frac{\nu_1(a^L(\zeta', z), \zeta')}{\nu_{11}(a^L(\zeta', z), \zeta')} - \frac{\nu_1(a(\zeta', z), \zeta')}{\nu_{11}(a(\zeta', z), \zeta')}\right] \\ &= -f'(z)\frac{\nu_1(a^L(\zeta', z), \zeta')}{\nu_{11}(a^L(\zeta', z), \zeta')}\end{aligned}$$

Since by concavity $\nu_1(a, \zeta') > 0$ for $a < a(\zeta', z)$ and $\nu_1(a, \zeta') < 0$ for $a > a(\zeta', z)$, we have $\frac{d\left|a^L(\zeta', z) - a(\zeta', z)\right|}{dz} \leq 0$. Similarly

$$\frac{dS(\zeta)}{dz} = -f'(z)\left[\frac{\nu_1(a^+(\zeta', z), \zeta')}{\nu_{11}(a^+(\zeta', z), \zeta')} - \frac{\nu_1(a^-(\zeta', z), \zeta')}{\nu_{11}(a^-(\zeta', z), \zeta')}\right]$$

Since $\nu_1(a^+(\zeta', z), \zeta') \leq 0$ and $\nu_1(a^-(\zeta', z), \zeta') \geq 0$, we have $\frac{dS(\zeta)}{dz} \leq 0$. Therefore Monotone Sensitivity holds and any set of environments that varies only in z is a valid comparison set.

Intuitively, this case captures changes in the slope of payoffs that leave the optimal natural action unchanged. For example, an increase in the scale of incentives that makes the payoff function "more concave" around the natural action makes deviating from the natural action more costly and so decreases the magnitude of latent demand and sensitivity.

Example 2 (Belief scoring)**.** *Consider a belief-reporting task rewarded by a quadratic scoring rule. A risk-neutral participant reports a belief, a, which is the probability of an event A. He is paid $\frac{z}{2}\left[1 - (\mathbb{I}[A] - a)^2\right]$ where $\mathbb{I}[A] = 1$ if A is true and 0 otherwise. The utility function is $U(a, \zeta', z) = \frac{z}{2}\left[1 - \mu(1-a)^2 - (1-\mu)(-a)^2\right] + a\phi(\zeta')E[h|\zeta']$, so $f(z) = z$. The optimal action solves $z\left[\mu(1-a^*) - (1-\mu)a^*\right] + \phi(\zeta')E[h|\zeta'] = 0$ or $a^* = \mu + \frac{\phi(\zeta')E[h|\zeta']}{z}$. Increases in z are equivalent to decreases in ϕ and decrease both the magnitude of latent demand effects, and sensitivity.*

Additive separability. Suppose that $v(a, \zeta', z) = v(a, \zeta') + af(z)$ and define z such that $f'(z) > 0$. Then:

$$\frac{d(a^L(\zeta', z) - a(\zeta', z))}{dz} = -f'(z)\left[\frac{1}{\nu_{11}(a^L(\zeta', z), \zeta')} - \frac{1}{\nu_{11}(a(\zeta', z), \zeta')}\right]$$

and

$$\frac{dS(\zeta)}{dz} = -f'(z)\left[\frac{1}{\nu_{11}(a^+(\zeta', z), \zeta')} - \frac{1}{\nu_{11}(a^-(\zeta', z), \zeta')}\right]$$

What matters in this case is how the concavity of v (and therefore ν) with respect to a varies with a. Suppose $\nu_{111} < 0$, so ν_{11} is decreasing in a, i.e. concavity is increasing. Then $\frac{dS(\zeta)}{dz} < 0$, i.e. increases in z decrease sensitivity. If $a^L(\zeta', z) > a(\zeta', z)$ then $\frac{d(a^L(\zeta', z) - a(\zeta', z))}{dz} < 0$ and if $a^L(\zeta', z) < a(\zeta', z)$ then $\frac{d(a^L(\zeta', z) - a(\zeta', z))}{dz} > 0$, so $\frac{d\left|a^L(\zeta', z) - a(\zeta', z)\right|}{dz} < 0$ and Monotone Sensitivity holds. Monotone Sensitivity also holds (with the inequalities reversed) for $\nu_{111} > 0$.

Example 3 (Effort provision)**.** *A participant performs a real-effort task for piece rate z with cost of effort $C(a)$, $C' > 0$,$C'' > 0$, $C''' > 0$. $U(a, \zeta', z) = za - C(a) + a\phi(\zeta')E[h|\zeta']$. The optimal action a^* solves $z - C'(a^*) + \phi(\zeta')E[h|\zeta'] = 0$. As z increases, a^* increases and responsiveness to latent demand or demand treatments decreases.*

Variation driven by inattention. Suppose that with some probability ξ the participant is an attentive type who pays careful attention to the decision-making environment, and with probability $1 - \xi$, he is inattentive. When inattentive, he takes some action $a^I(\zeta)$. $a^I(\zeta)$ might be equal to $a(\zeta)$, in which case the participant is only inattentive to experimenter demand, but it might differ if the participant is also inattentive to other design features.

While until now we have treated the actions as those of a representative agent, for this analysis it is more natural to work with expected or average actions over a sample. Denote by $\bar{a}(\zeta) = \xi a(\zeta) + (1-\xi)a^I(\zeta)$ the expected natural action, define $\bar{a}^L(\zeta), \bar{a}^+(\zeta), \bar{a}^-(\zeta)$ equivalently and let $\bar{S}(\zeta) = \bar{a}^+(\zeta) - \bar{a}^-(\zeta)$. The latent demand effect is now equal to $\left|\bar{a}^L(\zeta) - \bar{a}(\zeta)\right| = \xi\left|a^L(\zeta) - a(\zeta)\right|$, while $\bar{S}(\zeta) = \xi S(\zeta)$. Hence, if the variation in latent demand effects is driven by variation in attention, ξ, Monotone Sensitivity will hold, and any set of environments that varies only in participant attentiveness is a valid comparison set. Note that since we have assumed the participant is inattentive to both latent demand and the demand treatment, Bounding will hold if $p^T \geq p^L$ as before.

Variation driven by beliefs. Consider changes to the environment that influence behaviour only by altering participants' beliefs about the experimenter's objective, i.e. we consider variation in $h^L(\zeta)p^L(\zeta)$. Call this term H. $a(\zeta)$ is unaffected, so:

$$\frac{d(a^L(\zeta) - a(\zeta))}{dH} = -\frac{\phi(\zeta)}{v_{11}(a^L(\zeta), \zeta)} \geq 0$$

and therefore $\frac{d\left|a^L(\zeta)-a(\zeta)\right|}{dH} = -\frac{\phi(\zeta)}{v_{11}(a^L(\zeta),\zeta)} \times \text{sign}(a^L(\zeta) - a(\zeta)) = -\frac{\phi(\zeta)h^L(\zeta)}{v_{11}(a^L(\zeta),\zeta)}$ which is positive when $h^L(\zeta) = 1$ (because an increase in H means the participant's beliefs are shifting toward certainty that the experimenter wants a high action) and negative when $h^L(\zeta) = -1$ (because the participant is becoming more uncertain about the experimenter's wishes).

Next we turn to demand treatment effects. First we derive the response of

the participant's posterior:

$$\begin{aligned}\frac{d\frac{H+h^Tp^T}{1+Hh^Tp^T}}{dH} &= \frac{\left(1+Hh^Tp^T\right)-\left(H+h^Tp^T\right)h^Tp^T}{\left(1+Hh^Tp^T\right)^2} \\ &= \frac{1-\left(h^Tp^T\right)^2}{\left(1+Hh^Tp^T\right)^2} = \frac{1-p^{T2}}{\left(1+Hh^Tp^T\right)^2}\end{aligned}$$

So:

$$\frac{dS(\zeta)}{dH} = -\phi(\zeta)(1-p^{T2})\left[\frac{1}{\left(1+Hp^T\right)^2 v_{11}(a^+(\zeta),\zeta)} - \frac{1}{\left(1-Hp^T\right)^2 v_{11}(a^-(\zeta),\zeta)}\right]$$

The sign of this expression depends on the sign of H and how v_{11} changes with a. However, it is straightforward to see that Monotone Sensitivity *will not* hold in general, and in fact sensitivity will tend to be higher when latent demand is weaker. To see this, consider the simple case where v_{11} is constant. Then we have:

$$\begin{aligned}\frac{dS(\zeta)}{dH} &= -\frac{\phi(\zeta)(1-p^{T2})}{v_{11}}\left[\frac{\left(1-Hp^T\right)^2-\left(1+Hp^T\right)^2}{\left(1+Hp^T\right)^2\left(1-Hp^T\right)^2}\right] \\ &= -\frac{\phi(\zeta)(1-p^{T2})}{v_{11}}\left[\frac{-4Hp^T}{\left(1+Hp^T\right)^2\left(1-Hp^T\right)^2}\right]\end{aligned}$$

which is positive when $h^L = -1$ and negative when $h^L = 1$, i.e. it has the opposite sign to $\frac{d\left|a^L(\zeta)-a(\zeta)\right|}{dH}$. The reason is that as H approaches zero, the participant becomes more uncertain about the experimenter's wishes and is therefore very responsive to the new information in the demand treatments. Meanwhile as H approaches 1 or -1, the participant is very confident about the value of h. Although his confidence can be undermined by a demand treatment in the opposite direction, he responds little to a demand treatment that confirms his beliefs, so sensitivity is low.

2.B.4 Defiers

In this section we discuss defiance. We first derive a special case that illustrates how valid bounds can be obtained even when some participants defy the experimenter. Then we present three examples where defier behavior causes our key assumptions to break down.

Because our concern is with bounding rather than point identification, the method is able to tolerate some defiance. To illustrate, suppose that v is homogeneous across individuals, quadratic in a, and normalised such that $v_1(a,\zeta) = b(\zeta) - a$ where b is a constant. The natural action is equal to b for all individuals. Beliefs and ϕ are heterogeneous across individuals, indexed by i. For compactness, label the beliefs $H_i^L := h_i^L p_i^L$, $H_i^+ := (H_i^L + p^T)/(1 + H_i^L p^T)$ and $H_i^- := (H_i^L - p^T)/(1 - H_i^L p^T)$. Under our assumptions, the actions of interest are given by:

$$a_i^L = b + \phi_i H_i^L \qquad a_i^+ = b + \phi_i H_i^+ \qquad a_i^- = b + \phi_i H_i^-$$

Then, for Bounding to hold on average in the population, we require $E[\phi_i H_i^+] \geq 0 \geq E[\phi_i H_i^-]$, where expectations are over participants. If $p^T \geq p_i^L$ for all individuals, then $H_i^+ \geq 0 \geq H_i^-$, so the conditions are equivalent to weighted averages of ϕ having the correct sign, where the weights are the beliefs. A special case of interest is that where all individuals have identical H_i^+ and H_i^- (this is the case if latent demand (H_i^L) is the same for all individuals, or if $p^T = 1$). Then both conditions reduce to $E[\phi_i] \geq 0$, i.e. Bounding holds if the average participant is a complier.[36]

[36]For Monotonicity to hold on average we require $E[\phi_i(H_i^+ - H_i^L)] \geq 0 \geq E[\phi_i(H_i^L - H_i^-)]$. Since $H_i^+ - H_i^L > 0$ and $H_i^L - H_i^- < 0$, these conditions require that a weighted average of ϕ has the correct sign, where the weights are the belief *changes* induced by the

Now we provide three simple examples where defier behaviour causes our key assumptions to break down. First we show that it is possible for Bounding to hold without Monotonicity, second that it is possible for Monotonicity to hold without Bounding, and third that both can fail while retaining well-ordered bounds.

Let all decision makers share $v(a) = -a^2$, so the natural action $a = 0$. $2/3$ of the population are compliers with $\phi = \phi_C = 1$ and $1/3$ are defiers with $\phi = \phi_D = -1$. Latent demand signals are assumed common within complier/defier groups but different between compliers and defiers. They equal $H_C^L = h_C^L p_C^L$ and $H_D^L = h_D^L p_D^L$ respectively, with corresponding beliefs following the demand treatments equal to:

$$H_i^+ = \frac{H_i^L + p^T}{1 + H_i^L p^T} \qquad H_i^- = \frac{H_i^L - p^T}{1 - H_i^L p^T}$$

We retain the assumption of common p^T. Then the observed average actions under latent demand, positive and negative demand treatments are:

$$E[a^L] = \frac{1}{3}(2H_C^L - H_D^L) \qquad E[a^+] = \frac{1}{3}(2H_C^+ - H_D^+) \qquad E[a^-] = \frac{1}{3}(2H_C^- - H_D^-)$$

Our first example shows that Bounding can hold without Monotonicity. Thus a Monotonicity failure does not imply a failure of Bounding, but it is a warning sign of the presence of defiers.

Example 4 (Bounding without Monotonicity)**.** *Suppose* $H_C^L = 0.5$, $H_D^L = -0.5$ *and* $p^T = 1$. *Then* $E[a^L] = 0.5$, $E[a^+] = 1/3$ *and* $E[a^-] = -1/3$. *Therefore* $E[a^-] < a < E[a^+] < E[a^L]$.

demand treatments. Violations of Monotonicity or, in the extreme case, reversed bounds ($a^- > a^+$), are clear cause for concern. However it is possible for Monotonicity to hold on average while Bounding fails and vice versa.

Our second example shows that Bounding can fail while Monotonicity holds. This is possible in the basic model if $p^T < p^L$, but can also be caused by defiance.

Example 5 (Monotonicity without Bounding.)**.** *Suppose* $H_C^L = 0.5$, $H_D^L = -0.5$ *and* $p^T = 0.5$. *Then* $E[a^L] = 0.5$, $E[a^+] = \frac{8}{15}$ *and* $E[a^-] = \frac{4}{15}$. *Thus* $a < E[a^-] < E[a^L] < E[a^+]$.

Our third example shows that both Bounding and Monotonicity can fail, while still producing well-ordered bounds (i.e. $a^+ > a^-$).

Example 6 (No Bounding or Monotonicity.)**.** *Let* $H_C^L = 0.75$, $H_D^L = 0$, *and* $p^T = 0.75$. *Then* $E[a^L] = 0.5$, $E[a^+] = \frac{39}{100}$ *and* $E[a^-] = 1/4$. *Therefore* $a < E[a^-] < E[a^+] < E[a^L]$.

2.B.5 Extension: learning about ϕ

A possible interpretation of our demand treatments is that they signal not only the direction of the experimenter's objective, but the salience or intensity of her preference over objectives. For instance, "do me a favour" suggests that the choice is important. In this section we extend the model to incorporate this feature, allowing ϕ to depend upon a belief about the "importance" of the objective. We assume that the decision-maker responds more strongly to experimenter demand when they believe that complying with the objective it is more important, and that this belief depends both on latent demand and the demand treatments. Specifically, we now assume that the decision-maker's preferences are:

$$U(a, \zeta) = v(a, \zeta) + a\phi(\zeta)E[gh|\zeta]$$

where $g \in \{0, 1\}$ captures whether conforming to h is important (1) or unimportant (0) to the experimenter. ϕ remains the decision-maker's preference for pleasing the experimenter, which is now scaled by g, i.e. the decision-maker internalizes the perceived importance of the objective. We assume that g and h are believed independent (i.e. direction and importance are independent), so $E[gh|\zeta] = E[g|\zeta]E[h|\zeta]$. We also assume for simplicity is that the decision-maker's prior $E[g] = 0.5$.

Now, ζ contains two signals, $h^L(\zeta)$, defined as before, and $g^L(\zeta) \in \{0, 1\}$, where $E[g|g^L(\zeta)] = E[g|g^L(\zeta), \zeta]$ (i.e. g^L is a sufficient statistic). g^L is believed to equal g with probability $q^L(\zeta) < 1$ and pure independent noise otherwise. We show below that $E[g|g^L(\zeta)] = \frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right)$.

Similarly, a demand treatment is now two signals (h^T, g^T), where h^T is defined as before and $g^T \in \{0, 1, \emptyset\}$. $g^T = \emptyset$ corresponds to the case where no treatment is used, $g^T = 0$ signals to the participant that their action is not important to the experimenter, and $g^T = 1$ signals that it is.

Conditional on sending a demand treatment, g^T is believed to equal g with probability q^T and otherwise be pure noise independent of all other signals. We show below that the Bayesian posterior is:

$$E[g|g^T, g^L(\zeta)] = \frac{\frac{1}{2} + q^L(\zeta)\left(g^L(\zeta) - \frac{1}{2}\right) + q^T\left(g^T - \frac{1}{2}\right) + q^T q^L(\zeta)\left(\mathbb{I}[g^T = g^L(\zeta)] - \frac{1}{2}\right)}{1 + 2q^T q^L(\zeta)\left(\mathbb{I}[g^T = g^L(\zeta)] - \frac{1}{2}\right)}$$

We assume that g^T can be varied independently of h^T and will be held constant within a typical pair of positive and negative demand treatments.

For Bounding to hold, we now need:

$$\phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 0, h^L(\zeta)] \leq 0 \leq \phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 1, h^L(\zeta)]$$

Since $E[g|g^T, g^L(\zeta)] \geq 0$ our Bounding condition does *not* depend on how the demand treatments affect beliefs about g, all we require is $\phi(\zeta) \geq 0$ and $p^T \geq p^L(\zeta)$ as before.[37]

However, beliefs about g do affect the width of the bounds: sensitivity is increasing in $E[g|g^T, g^L(\zeta)]$. The tightest bounds are obtained when $E[g|g^T, g^L(\zeta)] = 0$, which obtains when $g^T = 0$ and $q^T = 1$. More generally, the bounds are tightened by signalling that acting according to the experimenter's objective is not important ($g^T = 0$), or if $g^T = 1$ by minimizing q^T. We suspect that it may be difficult in practice to both strongly signal the direction of the objective (large p^T), which is required for Bounding, and that the objective is not important ($g^T = 0$), so reasonable demand treatments are likely to be those that strongly signal a directional objective while keeping salience low, i.e. large p^T and small q^T with $g^T = 1$.

Derivation of $E[g|g^L(\zeta)]$ and $E[g|g^T, g^L(\zeta)]$ Let the prior belief be $\frac{1}{2}$.

$$E[g|g^L = y] = Pr(g = 1|g^L = y) = \frac{A}{B}$$
$$A = Pr(g^L = y|g = 1)Pr(g = 1)$$
$$B = Pr(g^L = y|g = 1)Pr(g = 1) + Pr(g^L = y|g = 0)Pr(g = 0)$$

[37] For Monotonicity to hold, we require

$$\phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 0, h^L(\zeta)] \leq \phi(\zeta)E[g|g^L(\zeta)]E[h|h^L(\zeta)] \leq \phi(\zeta)E[g|g^T, g^L(\zeta)]E[h|h^T = 1, h^L(\zeta)]$$

We can write

$$\phi(\zeta)\frac{E[h|h^T = 0, h^L(\zeta)]}{E[h|h^L(\zeta)]} \leq \phi(\zeta)\frac{E[g|g^L(\zeta)]}{E[g|g^T, g^L(\zeta)]} \leq \phi(\zeta)\frac{E[h|h^T = 1, h^L(\zeta)]}{E[h|h^L(\zeta)]}$$

We see that $\phi(\zeta) \geq 0$ is necessary but not sufficient for Monotonicity, we also need that $E[g|g^T, g^L(\zeta)]$ is neither "too big" nor "too small" relative to $E[g|g^L(\zeta)]$. Intuitively, if $g^T = 1$ the demand treatments shift all actions further away from the natural action, while if $g^T = 0$. all actions are shifted toward the natural action. $g^T = 1$ and $p^T \geq p^L$ are sufficient for Monotonicity to hold.

Since $Pr(g = j|g^L = y) = \frac{1}{2}(1 - q^L) + q^L\mathbb{I}[y = j]$ and $Pr(g = j) = \frac{1}{2}$ we have

$$\begin{aligned}
A &= \frac{1}{2}\left(\frac{1}{2}(1 - q^L) + q^L\mathbb{I}[y = 1]\right) \\
&= \frac{1}{2}\left(\frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right)\right) \\
B &= \frac{1}{2}\left[\left(\frac{1}{2}(1 - q^L) + q^L\mathbb{I}[y = 1]\right) + \left(\frac{1}{2}(1 - q^L) + q^L\mathbb{I}[y = 0]\right)\right] = \frac{1}{2}
\end{aligned}$$

Therefore, $E[g|g^L(\zeta)] = \frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right)$.

Turning to $E[g|g^T, g^L(\zeta)]$, we have assumed that when $g^T = \emptyset$, $E[g|g^T, g^L] = E[g|g^L]$. After observing $g^T \neq \emptyset$, the participant forms a posterior:

$$\begin{aligned}
E[g|g^T, g^L] = Pr(g = 1|g^T, g^L) &= \frac{A}{B} \\
A &= Pr(g^T = x|g = 1, g^L = y)Pr(g = 1|g^L = y) \\
B &= Pr(g^T = x|g = 1, g^L = y)Pr(g = 1|g^L = y) \\
&\quad + Pr(g^T = x|g = 0, g^L = y)Pr(g = 0|g^L = y)
\end{aligned}$$

Using the following

$$\begin{aligned}
Pr(g^T = x|g = j, g^L = y) &= \frac{1}{2}(1 - q^T) + q^T\mathbb{I}[x = j] \\
Pr(g = j|g^L = y) &= \frac{1}{2}(1 - q^L) + q^L\mathbb{I}[y = j]
\end{aligned}$$

we have:

$$
\begin{aligned}
A &= \left(\frac{1}{2}(1-q^T)+q^T\mathbb{I}[x=1]\right)\left(\frac{1}{2}(1-q^L)+q^L\mathbb{I}[y=1]\right)\\
&= \left(\frac{1}{2}(1-q^T)+q^Tg^T\right)\left(\frac{1}{2}(1-q^L)+q^Lg^L\right)\\
&= \frac{1}{2}q^Lg^L+\frac{1}{2}q^Tg^T-\frac{1}{2}q^Tq^L\left(g^L(1-g^T)+g^T(1-g^L)\right)\\
&+\frac{1}{4}(1-q^T)(1-q^L)\\
&= \frac{1}{2}q^Lg^L+\frac{1}{2}q^Tg^T-\frac{1}{2}q^Tq^L\left(\mathbb{I}[g^L\neq g^T]\right)\\
&+\frac{1}{4}-\frac{1}{4}q^T-\frac{1}{4}q^L+\frac{1}{4}q^Tq^L\\
&= \frac{1}{2}q^L\left(g^L-\frac{1}{2}\right)+\frac{1}{2}q^T\left(g^T-\frac{1}{2}\right)-\frac{1}{2}q^Tq^L\left(1-\mathbb{I}[g^L=g^T]\right)\\
&+\frac{1}{4}+\frac{1}{4}q^Tq^L\\
&= \frac{1}{4}+\frac{1}{2}q^L\left(g^L-\frac{1}{2}\right)+\frac{1}{2}q^T\left(g^T-\frac{1}{2}\right)\\
&+\frac{1}{2}q^Tq^L\left(\mathbb{I}[g^T=g^L]-\frac{1}{2}\right)
\end{aligned}
$$

$$
\begin{aligned}
B &= \left(\frac{1}{2}(1-q^T)+q^T\mathbb{I}[x=1]\right)\left(\frac{1}{2}(1-q^L)+q^L\mathbb{I}[y=1]\right)\\
&+\left(\frac{1}{2}(1-q^T)+q^T\mathbb{I}[x=0]\right)\left(\frac{1}{2}(1-q^L)+q^L\mathbb{I}[y=0]\right)\\
&= \left(\frac{1}{2}(1-q^T)+q^Tg^T\right)\left(\frac{1}{2}(1-q^L)+q^Lg^L\right)\\
&+\left(\frac{1}{2}(1-q^T)+q^T(1-g^T)\right)\left(\frac{1}{2}(1-q^L)+q^L(1-g^L)\right)\\
&= \frac{1}{2}(1-q^T)q^L+\frac{1}{2}(1-q^L)q^T+\frac{1}{2}(1-q^T)(1-q^L)\\
&+q^Tq^L\mathbb{I}[g^T=g^L]\\
&= \frac{1}{2}+q^Tq^L\left(\mathbb{I}[g^T=g^L]-\frac{1}{2}\right)
\end{aligned}
$$

Therefore,

$$E[g|g^T, g^L] = \frac{\frac{1}{2} + q^L\left(g^L - \frac{1}{2}\right) + q^T\left(g^T - \frac{1}{2}\right) + q^T q^L\left(\mathbb{I}[g^T = g^L] - \frac{1}{2}\right)}{1 + 2q^T q^L\left(\mathbb{I}[g^T = g^L] - \frac{1}{2}\right)}$$

2.B.6 Richer beliefs and correlated signals

Researchers sometimes give experimental participants instructions like "there are no right or wrong answers" or "we are only interested in what you think is the best choice". This can be thought of as a demand treatment that demands participants choose the natural action, $a(\zeta)$.

It is straightforward to analyze such treatments in our framework. In this section, we extend the model to allow h to take three values: $\{-1, 0, 1\}$, where $h = 0$ captures the case where the experimenter wants the participant to choose the natural action. We call the action following $h^T = 0$, $a^0(\zeta)$.

For simplicity we assume that the participant's prior belief is that each possibility is equally likely (i.e. is true with probability $1/3$), so $E[h] = 0$. ϵ and η are also believed to take each value with probability $1/3$ and are independent. $h^L \in \{-1, 0, 1\}$ and $h^T \in \{-1, 0, 1, \emptyset\}$ and p^L and p^T are defined as before. We maintain the assumption that the participant infers nothing when the experimenter does not send a demand treatment ($h^T = \emptyset$).

We show below that the beliefs can be written as:

$$E[h|h^L] = p^L h^L \tag{2.17}$$

$$E[h|h^T = \emptyset, h^L] = p^L h^L \tag{2.18}$$

$$E[h|h^T, h^L] = \frac{\frac{1}{3}(1-p^T)p^L h^L + \frac{1}{3}(1-p^L)p^T h^T + p^T p^L h^T \mathbb{I}[h^T = h^L]}{\frac{1}{3}\left(1 - p^T p^L\right) + p^T p^L \mathbb{I}[h^T = h^L]} \tag{2.19}$$

Bounding holds if $E[h|h^T = 1, h^L] \geq 0$ and $E[h|h^T = -1, h^L] \leq 0$. It is straightforward to check that the condition is the same as before: $p^T \geq p^L$.

What purpose, then, do $h^T = 0$ treatments serve? It is natural to think that demanding participants to take the natural action will eliminate demand effects, but under our assumptions, $h^T = 0$ does not in general elicit the natural action. Instead latent demand still influences the participant's action. We have:

$$E[h|h^T = 0, h^L] = \frac{\frac{1}{3}(1-p^T)p^L h^L}{\frac{1}{3}\left(1 - p^T p^L\right) + p^T p^L \mathbb{I}[h^L = 0]}$$

This expression equals zero if $p^T = 1$ (the demand treatment is perfectly informative), or $p^L h^L = 0$ (no latent demand), otherwise it has the same sign as $p^L h^L$. One interpretation is that while the participant takes at face value the experimenter's demand to choose the natural action, he might be unaware of the influence of other design features that nudge him in one direction or another.

Despite this negative result, $h^T = 0$ treatments can still be useful. First, they are informative about the *sign* of the bias due to latent demand. This is because $E[h|h^T = 0, h^L] \in [\min\{E[h|h^L], 0\}, \max\{E[h|h^L], 0\}]$ and therefore $a^0(\zeta) \in [\min\{a^L(\zeta), a(\zeta)\}, \max\{a^L(\zeta), a(\zeta)\}]$.[38] The action taken when $h^T = 0$

[38]To see this, note that $\left|E[h|h^L] - E[h|h^T = 0, h^L]\right| \geq 0$ and both have the same sign.

lies between the natural action and the action induced by latent demand, because the demand treatment shifts the participant's posterior toward zero.

Second, they can be used to obtain tighter bounds on $a(\zeta)$ if we know the direction of the latent demand effect. Suppose for example we know that $a^L(\zeta) \geq a(\zeta)$ (either from prior information or because we ran a treatment with $h^T = 0$ and verified that $a^0(\zeta) \leq a^L(\zeta)$). Then, the interval $[a^-(\zeta), a^0(\zeta)]$ gives a valid and tighter bound on $a(\zeta)$ than $[a^-(\zeta), a^+(\zeta)]$. Formally $a(\zeta) \in [a^-(\zeta), a^0(\zeta)] \subseteq [a^-(\zeta), a^+(\zeta)]$.[39]

Finally, there is one important case in which $h^T = 0$ perfectly recovers the natural action, i.e. $a^0(\zeta) = a(\zeta)$. Suppose that instead of assuming that the signals h^T and h^L contain independent shocks, the participant perceives that h^L is a noisy signal of h^T. Formally, he believes that with probability $p^L < 1$, $h^L = h^T$ and with probability $(1 - p^L)$, $h^L = \epsilon$. Then, when h^T and h^L disagree, he knows that h^L is pure noise, when they agree h^L contains no more information than h^T. Hence, the participant disregards h^L after observing h^T and $E[h|h^T, h^L] = p^T h^T$. Then, sending $h^T = 0$ recovers the natural action: $E[h|h^T = 0, h^L] = 0, \forall h^L$. An advantage of our bounds is that they are valid whether or not h^T or h^L are perceived as independent, in other words they are conservative relative to the approach of simply measuring $a^0(\zeta)$.

To summarize, unless the demand treatment is perceived as fully informative ($p^T = 1$), signalling $h^T = 0$ does *not* induce the participant to take the natural action, i.e. $a^0(\zeta) \neq a(\zeta)$. The intuition is that such a treatment does not eliminate all of the influence of latent demand – the decision-maker views both signals as informative and weighs them against one another,

[39] We thank Liad Weiss for pointing this out to us.

so the posterior belief lies between 0 and $E[h|h^L]$. However, because signalling $h^T = 0$ moves actions toward the natural action it can be informative about the *direction* of latent demand. In contrast, in an alternative formulation with non-independent signals, where participants perceive the demand treatments to contain the same information as latent demand but less noise, signalling $h^T = 0$ does elicit the natural action. Thus, demanding the natural action does not necessarily obtain bounds that contain the natural action, while a pair of sufficiently informative positive and negative demand treatments does.

Derivation of beliefs with ternary signals Recall that now $h \in \{-1, 0, 1\}$, $h^L \in \{-1, 0, 1\}$ and $h^T \in \{-1, 0, 1, \emptyset\}$.

To avoid clutter we suppress dependence on ζ. After observing h^L, the participant forms a posterior $E[h|h^L] = Pr(h = 1|h^L) \times 1 + Pr(h = -1|h^L) \times (-1)$. We can write this as:

$$
\begin{aligned}
E[h|h^L = y] = Pr(h = 1|h^L = y) - Pr(h = -1|h^L = y) = \frac{A}{B} \\
A = Pr(h^L = y|h = 1)Pr(h = 1) - Pr(h^L = y|h = -1)Pr(h = -1) \\
B = Pr(h^L = y|h = 1)Pr(h = 1) + Pr(h^L = y|h = 0)Pr(h = 0) \\
+ Pr(h^L = y|h = -1)Pr(h = -1)
\end{aligned}
$$

Since $Pr(h = j|h^L = y) = \frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = j]$ and $Pr(h = j) = \frac{1}{3}$ we have

$$\begin{aligned}
A &= \frac{1}{3}\left[\left(\frac{1}{3}(1-p^L) + p^L\mathbb{I}[y=1]\right) - \left(\frac{1}{3}(1-p^L) + p^L\mathbb{I}[y=-1]\right)\right] \\
&= \frac{1}{3}p^L\left[\mathbb{I}[y=1] - \mathbb{I}[y=-1]\right] = \frac{1}{3}p^L h^L \\
B &= \frac{1}{3}\left[\left(\frac{1}{3}(1-p^L) + p^L\mathbb{I}[y=1]\right) + \left(\frac{1}{3}(1-p^L) + p^L\mathbb{I}[y=0]\right)\right. \\
&\quad \left. + \left(\frac{1}{3}(1-p^L) + p^L\mathbb{I}[y=-1]\right)\right] = \frac{1}{3}
\end{aligned}$$

So

$$E[h|h^L = y] = p^L h^L \tag{2.20}$$

just as before. Turning to beliefs following the demand treatments, as before we assume that when $h^T = \emptyset$, $E[h|h^T, h^L] = E[h|h^L]$. We have:

$$\begin{aligned}
E[h|h^T, h^L] &= Pr(h = 1|h^T, h^L) - Pr(h = -1|h^T, h^L) = \frac{A}{B} \\
A &= Pr(h^T = x|h = 1, h^L = y)Pr(h = 1|h^L = y) \\
&\quad - Pr(h^T = x|h = -1, h^L = y)Pr(h = -1|h^L = y) \\
B &= Pr(h^T = x|h = 1, h^L = y)Pr(h = 1|h^L = y) \\
&\quad + Pr(h^T = x|h = 0, h^L = y)Pr(h = 0|h^L = y) \\
&\quad + Pr(h^T = x|h = -1, h^L = y)Pr(h = -1|h^L = y, h^L = y)
\end{aligned}$$

Using

$$\begin{aligned}
Pr(h^T = x|h = j, h^L = y) &= \frac{1}{3}(1 - p^T) + p^T\mathbb{I}[x = j] \\
Pr(h = j|h^L = y) &= \frac{1}{3}(1 - p^L) + p^L\mathbb{I}[y = j]
\end{aligned}$$

we obtain:

$$
\begin{aligned}
A &= \left(\frac{1}{3}(1-p^T)+p^T\mathbb{I}[x=1]\right)\left(\frac{1}{3}(1-p^L)+p^L\mathbb{I}[y=1]\right) \\
&- \left(\frac{1}{3}(1-p^T)+p^T\mathbb{I}[x=-1]\right)\left(\frac{1}{3}(1-p^L)+p^L\mathbb{I}[y=-1]\right) \\
&= \frac{1}{3}(1-p^T)p^L\mathbb{I}[y=1]+\frac{1}{3}(1-p^L)p^T\mathbb{I}[x=1]+p^Tp^L\mathbb{I}[x=1]\mathbb{I}[y=1] \\
&- \frac{1}{3}(1-p^T)p^L\mathbb{I}[y=-1]-\frac{1}{3}(1-p^L)p^T\mathbb{I}[x=-1]-p^Tp^L\mathbb{I}[x=-1]\mathbb{I}[y=-1] \\
&= \frac{1}{3}(1-p^T)p^Lh^L+\frac{1}{3}(1-p^L)p^Th^T+p^Tp^Lh^T\mathbb{I}[h^T=h^L]
\end{aligned}
$$

$$
\begin{aligned}
B &= \left(\frac{1}{3}(1-p^T)+p^T\mathbb{I}[x=1]\right)\left(\frac{1}{3}(1-p^L)+p^L\mathbb{I}[y=1]\right) \\
&+ \left(\frac{1}{3}(1-p^T)+p^T\mathbb{I}[x=0]\right)\left(\frac{1}{3}(1-p^L)+p^L\mathbb{I}[y=0]\right) \\
&+ \left(\frac{1}{3}(1-p^T)+p^T\mathbb{I}[x=-1]\right)\left(\frac{1}{3}(1-p^L)+p^L\mathbb{I}[y=-1]\right) \\
&= \frac{1}{3}(1-p^T)(1-p^L)+\frac{1}{3}p^T(1-p^L)\left(\mathbb{I}[x=1]+\mathbb{I}[x=0]+\mathbb{I}[x=-1]\right) \\
&+ \frac{1}{3}p^L(1-p^T)\left(\mathbb{I}[y=1]+\mathbb{I}[y=0]+\mathbb{I}[y=-1]\right) \\
&+ p^Tp^L\left(\mathbb{I}[x=1]\mathbb{I}[y=1]+\mathbb{I}[x=0]\mathbb{I}[y=0]+\mathbb{I}[x=-1]\mathbb{I}[y=-1]\right) \\
&= \frac{1}{3}\left(1-p^Tp^L\right)+p^Tp^L\mathbb{I}[h^T=h^L]
\end{aligned}
$$

So

$$
E[h|h^T,h^L]=\frac{\frac{1}{3}(1-p^T)p^Lh^L+\frac{1}{3}(1-p^L)p^Th^T+p^Tp^Lh^T\mathbb{I}[h^T=h^L]}{\frac{1}{3}\left(1-p^Tp^L\right)+p^Tp^L\mathbb{I}[h^T=h^L]} \tag{2.21}
$$

2.B.7 Computing confidence intervals

Here we describe how we compute demand-robust confidence intervals. We note that this was not included in the pre-analysis plans.

Confidence intervals for actions Imbens and Manski (2004) show that asymptotically the probability that the estimate for the upper (lower) bound is lower (higher) than the true value can be ignored when making inference. Thus, one can construct one-sided intervals with confidence level α around both the upper and the lower bound. The 95 percent confidence interval for the true demand-free behaviour is thus given by:

$$CI() = [a^-(\zeta) - \overline{C_N}\frac{\widehat{\sigma^-}}{\sqrt{N}},\ a^+(\zeta) + \overline{C_N}\frac{\widehat{\sigma^+}}{\sqrt{N}}]$$

Here, $\widehat{\sigma^-} = \sqrt{\widehat{Var(a^-(\zeta))}}$ and $\widehat{\sigma^+} = \sqrt{\widehat{Var(a^+(\zeta))}}$, and $\overline{C_N}$ satisfies

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{a^+(\zeta) - a^-(\zeta)}{\max(\widehat{\sigma^-},\ \widehat{\sigma^+})}\right) - \Phi(-\overline{C_N}) = 0.90.$$

The 95 percent confidence interval for the set $[a^-(\zeta), a^+(\zeta)]$ is given by:

$$CI() = [a^-(\zeta) - \overline{C_N}\frac{\widehat{\sigma^-}}{\sqrt{N}},\ a^+(\zeta) + \overline{C_N}\frac{\widehat{\sigma^+}}{\sqrt{N}}],$$

where $\overline{C_N}$ satisfies

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{a^+(\zeta) - a^-(\zeta)}{\max(\widehat{\sigma^-},\, \widehat{\sigma^+})}\right) - \Phi(-\overline{C_N}) = 0.95.$$

Confidence intervals for treatment effects We also outline how one can compute confidence intervals for the treatment effects $[a(\zeta_1) - a(\zeta_0)]$ and for the set defined by the upper and lower bounds for treatment effects as given by our demand treatments: $[a(\zeta_1) - a(\zeta_0)] \in [a^-(\zeta_1) - a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)]$

For simplicity we denote the lower bound, $[a^-(\zeta_1) - a^+(\zeta_0)]$, as T^- and the upper bound, $[a^+(\zeta_1) - a^-(\zeta_0)]$, as T^+. The 95 percent confidence interval for the true demand-free treatment effect is given by:

$$CI() = [T^- - \overline{C_N}\frac{\widehat{\sigma^{T-}}}{\sqrt{N}},\, T^+ + \overline{C_N}\frac{\widehat{\sigma^{T+}}}{\sqrt{N}}].$$

Here, $\widehat{\sigma^{T-}} = \sqrt{\widehat{Var(T^-)}}$ and $\widehat{\sigma^{T+}} = \sqrt{\widehat{Var(T^+)}}$, and $\overline{C_N}$ satisfies

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{T^+ - T^-}{\max(\widehat{\sigma^{T-}},\, \widehat{\sigma^{T+}})}\right) - \Phi(-\overline{C_N}) = 0.90.$$

The 95 percent confidence interval for the set $[a^-(\zeta_1) - a^+(\zeta_0), a^+(\zeta_1) - a^-(\zeta_0)]$ is as follows:

$$CI() = [T^- - \overline{C_N}\frac{\widehat{\sigma^{T-}}}{\sqrt{N}},\, T^+ + \overline{C_N}\frac{\widehat{\sigma^{T+}}}{\sqrt{N}}],$$

where

$$\Phi\left(\overline{C_N} + \sqrt{N}\frac{T^+(\tau) - T^-(\tau)}{\max(\widehat{\sigma^{T-}}, \widehat{\sigma^{T+}}}\right) - \Phi(-\overline{C_N}) = 0.95.$$

2.B.8 Controlling for demand

Here we provide derivations for the results in section 2.4.4. We begin with the usual first-order condition, assuming a demand treatment h^T:

$$0 = v_1(a^*(\zeta, h^T), \zeta) + \phi(\zeta)E[h|h^T, h^L(\zeta)]$$

taking the first-order Taylor approximation at the natural action $a(\zeta)$ we obtain:

$$0 \approx \underbrace{v_1(a(\zeta), \zeta)}_{=0} + \phi(\zeta)E[h|h^T, h^L(\zeta)] + [a^*(\zeta, h^T) - a(\zeta)]v_{11}(a(\zeta), \zeta)$$

where the first term is zero by definition of $a(\zeta)$. Rearranging we obtain equation (2.8):

$$\begin{aligned} a^*(\zeta, h^T) &\approx a(\zeta) - \frac{\phi(\zeta)}{v_{11}(a(\zeta), \zeta)}E[h|h^T, h^L(\zeta)]. \\ &= a(\zeta) + \Phi(\zeta)E[h|h^T, h^L(\zeta)]. \end{aligned}$$

Assume two treatment groups: $\zeta \in \{0, 1\}$, and denote their corresponding demand treatments by h^T_ζ. If no demand treatments are applied ($h^T_1 = h^T_0 = \emptyset$). Since we are interested in cases where $h^T_0 = h^T_1$ we suppress the sub-

scripts. The approximate bias of the treatment effect estimate can be written as:

$$
\begin{aligned}
Bias &= a^*(1,\emptyset) - a^*(0,\emptyset) - [a(1) - a(0)] \\
&\approx a(1) + \Phi(1)E[h|h^T, h^L(1)] - a(0) - \Phi(0)E[h|h^T, h^L(0)] - [a(1) - a(0)]
\end{aligned}
$$

Adding and subtracting terms yields:

$$
Bias \approx \underbrace{\Phi(1)\left(E[h|h^T, h^L(1)] - E[h|h^T, h^L(0)]\right)}_{\text{Bias due to beliefs}} + \underbrace{\left(\Phi(1) - \Phi(0)\right)E[h|h^T, h^L(0)]}_{\text{Bias due to ``responsiveness''}}
$$

Fully informative demand treatments We now show that when demand treatments are fully informative, one can test for bias due to behavioural responsiveness.

$$
\begin{aligned}
&\underbrace{[a^*(1,1) - a^*(1,0)]}_{\text{Sensitivity } (\zeta = 1)} - \underbrace{[a^*(0,1) - a^*(0,0)]}_{\text{Sensitivity } (\zeta = 0)} \\
&= \underbrace{[a^*(1,1) - a^*(0,1)]}_{\text{Treatment effect } (h^T = 1)} - \underbrace{[a^*(1,0) - a^*(0,0)]}_{\text{Treatment effect } (h^T = -1)} \\
&\approx \left[a(1) + \Phi(1)E[h|1, h^L(1)] - a(0) - \Phi(0)E[h|1, h^L(0)]\right] \\
&- \left[a(1) + \Phi(1)E[h|-1, h^L(1)] - a(0) - \Phi(0)E[h|-1, h^L(0)]\right] \\
&= [\Phi(1) \times 1 - \Phi(0) \times 1] - [\Phi(1) \times -1 - \Phi(0) \times -1] \\
&= 2\left(\Phi(1) - \Phi(0)\right)
\end{aligned}
$$

Next, we show that averaging the "positive-positive" and "negative-negative"

treatment effects approximates the true treatment effect

$$\begin{aligned}
&\frac{1}{2}\left([a^*(1,1)-a^*(0,1)]+[a^*(1,-1)-a^*(0,-1)]\right)\\
\approx&a(1)-a(0)+\frac{1}{2}\left[(\Phi(1)-\Phi(0))\times 1+(\Phi(1)-\Phi(0))\times -1\right]\\
=&a(1)-a(0)
\end{aligned}$$

Less informative treatments For compactness we define notation $H^L(\zeta) \equiv h^L(\zeta)p^L(\zeta)$ and $H^T \equiv h^T p^T$. Our Bounding assumption implies $\left|H^T\right| \geq \left|H^L(\zeta)\right|$.

Consider the expressions for belief differences between treatment and control, first without (Diff^L) and then with (Diff^T) demand treatments. We have

$$\text{Diff}^L \equiv H^L(1) - H^L(0)$$

and:

$$\begin{aligned}
\text{Diff}^T &\equiv \frac{H^L(1)+H^T}{1+H^L(1)H^T} - \frac{H^L(0)+H^T}{1+H^L(0)H^T}\\
&= \frac{(1-H^{T2})}{(1+H^L(1)H^T)(1+H^L(0)H^T)} \times \text{Diff}^L
\end{aligned}$$

We want to find conditions under which $\left|\text{Diff}^T\right| < \left|\text{Diff}^L\right|$, which holds if and only if

$$\frac{(1-H^{T2})}{(1+H^L(1)H^T)(1+H^L(0)H^T)} < 1$$

rearranging we obtain:

$$0 < H^T(H^L(1)+H^L(0)) + H^{T2}(1+H^L(1)H^L(0)) \tag{2.22}$$

If $h^T = 1$, (2.22) reduces to

$$-\frac{H^L(1) + H^L(0)}{1 + H^L(1)H^L(0)} < H^T$$

while if $h^T = -1$ it reduces to

$$-\frac{H^L(1) + H^L(0)}{1 + H^L(1)H^L(0)} > H^T.$$

Since the left hand side lies in the interval $(-1, 1)$ there always exists a sufficiently strong demand treatment (p^T sufficiently large) that (2.22) is satisfied. We now evaluate whether there is more we can say. There are X cases to consider. Assume throughout, without loss of generality, that $\left|H^L(1)\right| > \left|H^L(0)\right|$.

1. Suppose the latent demand beliefs have the same sign as each other ($h^L(1) = h^L(0) = h^L$) and the same sign as the demand treatment ($h^L = h^T$). Then it is easy to verify that (2.22) holds for all H^T. Intuitively, when the latent demand beliefs have the same sign, additional information that further reinforces those beliefs has a greater effect on the one that is less certain, reducing the gap between them.

2. The latent demand beliefs have the same sign as each other ($h^L(1) = h^L(0) = h^L$) and the opposite sign to the demand treatment ($h^L = -h^T$). Assume $h^T = 1$ and $h^L = -1$ (the opposite case is symmetric). We know that (2.22) holds for sufficiently strong H^T, we will ask if our Bounding assumption is sufficient. We show that it is not, by contradiction. Suppose Bounding holds exactly, i.e. $H^T = -H^L(1)$. Then, by

the premise that (2.22) is satisfied:

$$-\frac{H^L(1)+H^L(0)}{1+H^L(1)H^L(0)} < -H^L(1)$$
$$\frac{1+\frac{H^L(0)}{H^L(1)}}{1+H^L(1)H^L(0)} < 1$$

which holds if and only if:

$$\frac{H^L(0)}{H^L(1)} < H^L(1)H^L(0)$$
$$1 < H^L(1)^2$$

a contradiction since $H^L(1) < 1$. Thus in this case the condition for demand treatments to reduce bias is stronger than Bounding.

3. The latent demand beliefs have opposite signs ($h^L(1) = -h^L(0)$), and the stronger belief ($H^L(1)$) has the same sign as h^T, i.e. $h^T = h^L(1)$. Focus again on the case where $h^T = 1$ (the opposite case is symmetric). We require:
$$-\frac{H^L(1)+H^L(0)}{1+H^L(1)H^L(0)} < H^T$$
It is easy to see that the condition is always satisfied since the left-hand side is negative.

4. The latent demand beliefs have opposite signs ($h^L(1) = -h^L(0)$), and the stronger belief has the opposite sign to h^T, i.e. $h^T = -h^L(1)$. Focus again on the case where $h^T = 1$ (the opposite case is symmetric). We know that (2.22) holds for sufficiently strong H^T, we will ask if our Bounding assumption is sufficient. Thus let $H^T = -H^L(1)$ (bounding

holds exactly). We require:

$$
\begin{aligned}
-\frac{H^L(1)+H^L(0)}{1+H^L(1)H^L(0)} &< -H^L(1) \\
\frac{1+\frac{H^L(0)}{H^L(1)}}{1+H^L(1)H^L(0)} &< 1 \\
\frac{H^L(0)}{H^L(1)} &< H^L(1)H^L(0) \\
1 &> H^L(1)^2
\end{aligned}
$$

which is satisfied. Thus Bounding is sufficient for (2.22) to hold.

2.B.9 Structural estimation

This section outlines step by step how the parameters are constructed in our NLLS estimation of the structural model in section 2.4.5.

Data and parameter adjustments First, we follow DP exactly in rounding effort scores to the nearest 100 (except for those in range $[1, 49]$ which we round to 25). This is because incentives were paid per 100 points, and we wish to avoid modeling effort choices that lie between two 100 point thresholds. We refer the reader to DP for further details.

Second, we make a couple of adjustments pre and post-estimation. First, we divide the rounded scores by 100. In other words, if effort a is measured in points, we compute $a' = a/100$ which is measured in hundreds of points. Second, we multiply the incentive, ζ, which is measured in cents per point, by 100 to express it as $\zeta' = 100\zeta$ which is measured in cents per 100 points. These transformations were helpful in achieving convergence of the estimator, which otherwise occasionally suffered from underflow problems.

However they change the interpretation of the parameters. Specifically, the intrinsic motivation parameter s and the preference for pleasing the experimenter, ϕ, will both be measured in units equivalent to cents per 100 points, while the cost function parameters will be expressed for effort measured in hundreds of points.

To aid comparability with DP we therefore re-transform the parameters after estimation. DP present their estimates of incentive parameters (which in our case are s and ϕ) in the same units, cents per 100 points, so we do not need to correct them. k and γ are reported for effort measured in points, so we transform our estimates for comparability. We derive the adjustments as follows. First, for the power cost function, we have:

$$U = (s + \zeta + \phi E[h|h^T, h^L])a - \frac{ka^{1+\gamma}}{1+\gamma}$$

Let $a' = \frac{a}{100}$ and $\zeta' = 100\zeta$. Then:

$$\begin{aligned} U &= \left(s + \frac{\zeta'}{100} + \phi E[h|h^T, h^L]\right) 100a' - \frac{k(100a')^{1+\gamma}}{1+\gamma} \\ &= \left(100s + \zeta' + 100\phi E[h|h^T, h^L]\right) a' - \frac{k(100a')^{1+\gamma}}{1+\gamma} \end{aligned}$$

giving rise to first-order condition:

$$\begin{aligned} 0 &= \left(100s + \zeta' + 100\phi E[h|h^T, h^L]\right) - ka'^{\gamma} 100^{1+\gamma} \\ a' &= \left(\frac{100s + \zeta' + 100\phi E[h|h^T, h^L]}{k100^{1+\gamma}}\right)^{\frac{1}{\gamma}} \\ \log(a') &= \frac{1}{\gamma} \log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{k^*}\right) \end{aligned}$$

where $s^* = 100s$, $\phi^* = 100\phi$ and $k^* = 100^{1+\gamma}k$. We leave s^* and ϕ^*, (which are in equivalent units to cents per 100 points) untransformed for compara-

bility with DP. In the tables we report $k = k^*/100^{1+\gamma}$ and its standard error, computed via the delta method.

For the exponential cost function we have:

$$\begin{aligned} U &= (s + \zeta + \phi E[h|h^T, h^L])a - \frac{k}{\gamma}\exp(\gamma a) \\ &= (s^* + \zeta' + \phi^* E[h|h^T, h^L])a' - \frac{k}{\gamma}\exp(100\gamma a') \end{aligned}$$

implying first-order condition:

$$\begin{aligned} 0 &= s^* + \zeta' + \phi^* E[h|h^T, h^L] - 100k\exp(100\gamma a') \\ a' &= \frac{1}{100\gamma}\log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{100k}\right) \\ &= \frac{1}{\gamma^*}\log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{k^*}\right) \end{aligned}$$

where $s^* = 100s$, and $\phi^* = 100\phi$ as before, while $\gamma^* = 100\gamma$, $k^* = 100k$. In the tables we report $\gamma = \gamma^*/100$ and $k = k^*/100$.

Error term To allow for the observed heterogeneity in effort, we follow DP in assuming heterogeneous effort costs, as follows. Let the cost of effort under power utility equal $ka^{1+\gamma}(1+\gamma)^{-1}\exp(-\gamma\epsilon)$ where $\epsilon \sim N(0, \sigma_\epsilon^2)$. Then our FOC becomes

$$\begin{aligned} 0 &= \left(100s + \zeta' + 100\phi E[h|h^T, h^L]\right) - ka'^{\gamma}100^{1+\gamma}\exp(-\gamma\epsilon) \\ a' &= \left(\frac{100s + \zeta' + 100\phi E[h|h^T, h^L]}{k100^{1+\gamma}}\right)^{\frac{1}{\gamma}}\exp(\epsilon) \\ \log(a') &= \frac{1}{\gamma}\log\left(\frac{100s + \zeta' + 100\phi E[h|h^T, h^L]}{k100^{1+\gamma}}\right) + \epsilon \end{aligned}$$

where ϵ becomes the error term in our NLLS routine. For the exponential cost, we follow DP and assume effort cost is $k\gamma^{-1}\exp(\gamma a)\exp(-\gamma\epsilon)$. Then our FOC becomes

$$\begin{aligned} 0 &= s^* + \zeta' + \phi^* E[h|h^T, h^L] - 100k\exp(100\gamma a')\exp(-\gamma\epsilon) \\ a' &= \frac{1}{100\gamma}\log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{100k}\right) + \frac{\epsilon}{100} \\ &= \frac{1}{\gamma^*}\log\left(\frac{s^* + \zeta' + \phi^* E[h|h^T, h^L]}{k^*}\right) + \epsilon^* \end{aligned}$$

where $\epsilon^* = \epsilon/100$ forms the error term in our estimation.

Estimating equation Finally, in our estimation we sometimes need to estimate the product $\phi^* E[h|h^L]$. We estimate this product directly, then transform by dividing by ϕ^*. Specifically, we estimate the following:

$$\begin{aligned} y_i = \frac{1}{\beta_0}\log\big[&\zeta_i' + \beta_1 + \beta_2(\text{pos_demand}_i - \text{neg_demand}_i) \\ &+ \beta_3 \times \text{no_demand}_i \times \text{incentive_0c}_i + \beta_4 \times \text{no_demand}_i \times \text{incentive_1c}_i \\ &+ \beta_5 \times \text{no_demand}_i \times \text{incentive_4c}_i\big] - \frac{1}{\beta_0}\log(\beta_6) + \varepsilon_i \end{aligned}$$

where $y = \log(a')$ or a' respectively, pos_demand, neg_demand and no_demand are dummies for our positive, negative and no demand treatments, while incentive_Xc is a dummy for the treatment with X cents per 100 points. Parameters are as follows: $\beta_0 = \gamma$ or γ^* respectively, $\beta_1 = s^*$, $\beta_2 = \phi^*$, $\beta_3 = \phi^* E[h|h^L(\zeta = 0)]$, $\beta_4 = \phi^* E[h|h^L(\zeta = 1)]$, $\beta_5 = \phi^* E[h|h^L(\zeta = 4)]$ and $\beta_6 = k^*$. We then compute the three values for $E[h|h^L]$ by dividing by β_2, i.e. β_3/β_2, β_4/β_2 and β_5/β_2. γ and k are computed by the transformations outlined above. Standard errors are computed by the delta method. In the specification where we restrict latent demand to be equal for the 1 cent and

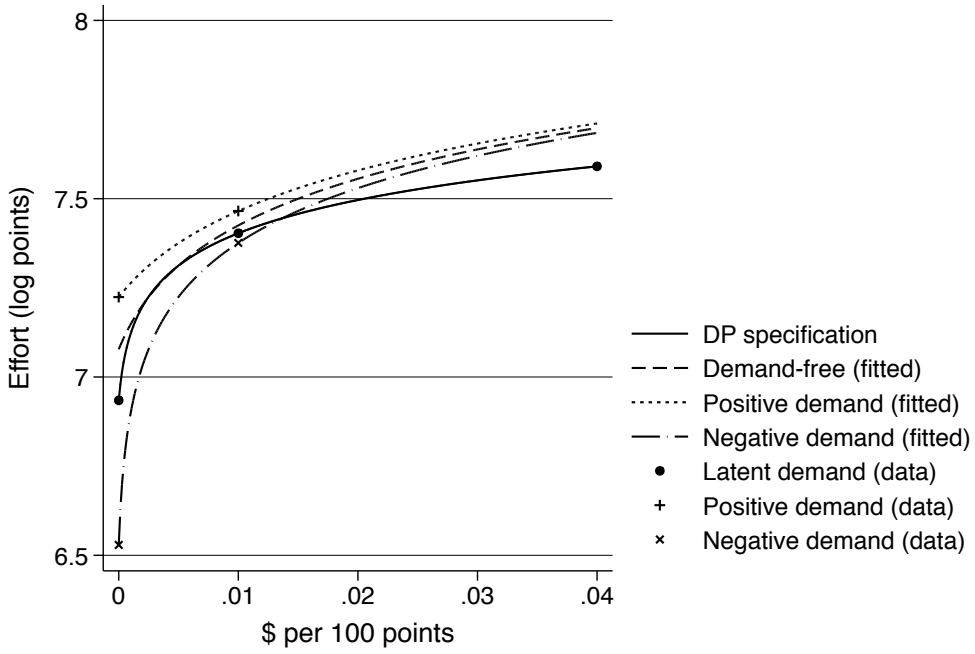
4 cent treatments we impose $\beta_4 = \beta_5$.

Extrapolation Our large estimates of $h^L(4)p^L(4)$ reflect an out-of-sample extrapolation as the main model parameters are estimated from the 0 and 1 cent treatment groups. Figure A.4 illustrates this for the power cost case. Points correspond to mean (log) effort for each treatment group. The figure then plots (a) predicted effort using the DP specification, which fits the "no demand" data only (parameters taken from Column 1 of table 2.6), and (b) predicted effort using the exactly identified model (Column 3 of table 2.6), for each case of zero demand $E[h] = 0$, strong positive demand ($E[h] = 1$) and strong negative demand ($E[h] = -1$). The estimation then recovers the latent demand beliefs by comparing observed effort to predicted effort when demand is zero.

It is clear from the figure that the extrapolation from model (b) to the 4 cent effort treatment is not perfect, and the observed behavior lies outside the limits implied by $E[h] \in [-1, 1]$. This is the reason for the large negative fitted value for beliefs at this point.

Another thing that is clear from the figure is how the curvature of the effort cost function determines the imputed latent beliefs, which may explain the difference in imputed beliefs between the power and exponential cost functions. The sign of imputed beliefs depends on whether the "no demand" point lies above or below the curve, so changes in curvature can flip the sign of these estimates.

Figure A.4: Structural estimation: fitted values



Notes: Figure displays mean (log) effort for each treatment used in the structural estimation, and fitted values from the estimated models.

2.B.10 Using the method in practice

In this section we work through an example to illustrate the methods we have developed. We do this with the help of Figure A.5, which represents the 10 data points available to a researcher who has applied no demand, weak and strong treatments to a control group (ζ_0) and a treatment group (ζ_1). To avoid confusion, we will label the actions under weak positive and negative demand as $a^+(\zeta)$ and $a^-(\zeta)$, while strong (assumed to be fully informative) positive and negative demand actions are defined as $a^{++}(\zeta)$ and $a^{--}(\zeta)$. In our example the conventional treatment effect is given by $a^L(\zeta_1) - a^L(\zeta_0)$ (or point H minus point C).

1. We can use our strong treatments to construct bounds for actions: In Figure A.5 the bounds on action one are defined by points A and E for the control group, $[a^{--}(\zeta_0), a^{++}(\zeta_0)]$, and by points F and J for the treatment group, $[a^{--}(\zeta_1), a^{++}(\zeta_1)]$.

2. Similarly, we can construct bounds using the weak treatments, which are defined by points B and D for the control group, $[a^-(\zeta_0), a^+(\zeta_0)]$, and by points G and I for the treatment group, $[a^-(\zeta_1), a^+(\zeta_1)]$. Under the assumption that our demand treatments are more informative than underlying latent demand, these bounds contain the natural action.

3. We can analogously also define strong bounds for treatment effects. The upper bound is given by the comparison between respondents in the treatment group that receive strong positive demand treatments, and respondents in the control group that receive strong negative demand treatments. In Figure A.5 this corresponds to the difference between points J and A: $a^{++}(\zeta_1) - a^{--}(\zeta_0)$. The lower bound is given

by the comparison between respondents in the treatment group that receive strong negative demand treatments, and respondents in the control group that receive strong positive demand treatments, given by the difference points F and E: $[a^{--}(\zeta_1) - a^{++}(\zeta_0)$. The bounds are formally defined as follows: $[a^{--}(\zeta_1) - a^{++}(\zeta_0), a^{++}(\zeta_1) - a^{--}(\zeta_0)]$.

4. Similarly, we can construct weak bounds for treatment effects, by applying weak instead of strong treatments. The bounds are given by: $[a^{-}(\zeta_1) - a^{+}(\zeta_0), a^{+}(\zeta_1) - a^{-}(\zeta_0)]$. In Figure A.5 the upper and lower bounds are given by the difference between points I and B, and points G and D, respectively.

5. An alternative to creating bounds, is to "control for demand effects". Under the assumption that demand treatments are fully informative, provided responsiveness to demand treatments does not differ across treatment groups, we can point identify treatment effects that are not biased by demand effects. In this specific case, we could apply strong positive or strong negative demand treatments to both the treatment and the control group: In Figure A.5 the estimates are given by the difference between points J and E: $(a^{++}(\zeta_1) - a^{++}(\zeta_0))$, or F and A: $(a^{--}(\zeta_1) - a^{--}(\zeta_0))$.

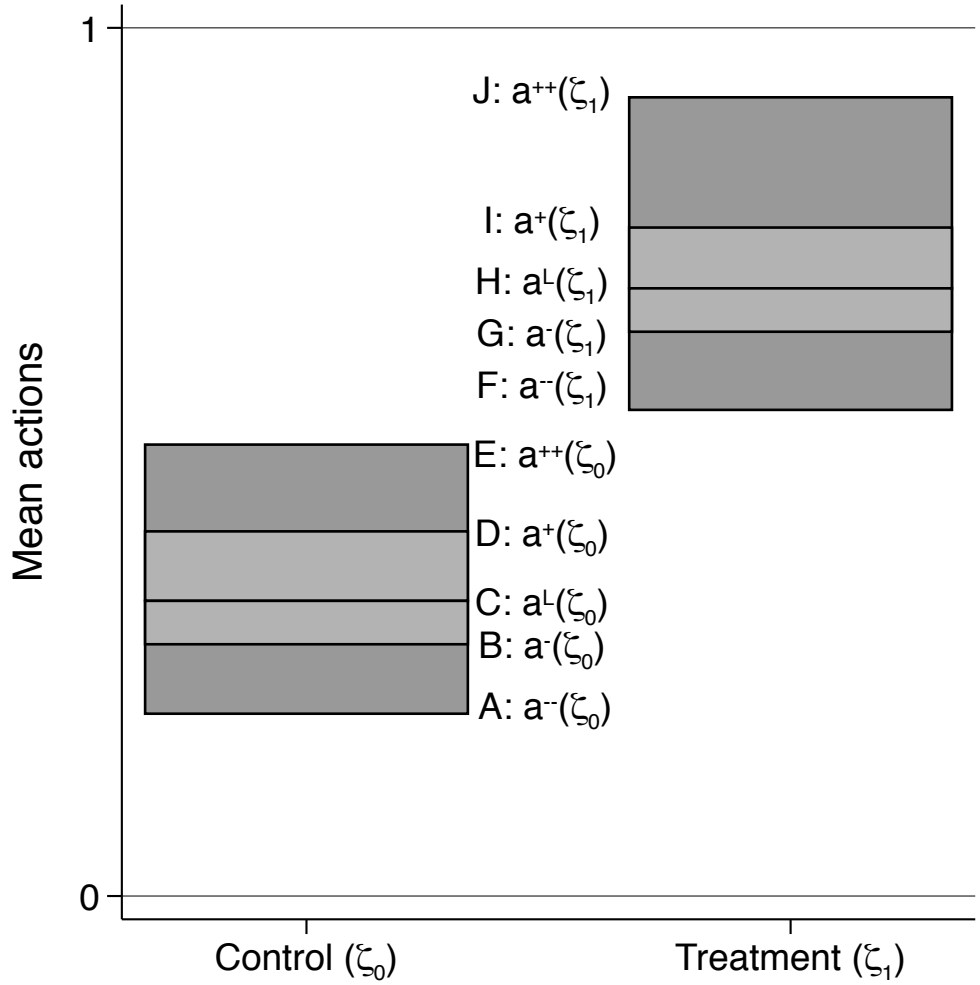
6. If responsiveness to fully informative demand treatments differs significantly across treatment arms, our point estimates from employing same-signed demand treatments are still biased. However, by the symmetry of the Taylor approximation, we can approximate the treatment effect using the midpoints of the bounds generated the strong demand treatments. In A.5 this corresponds to comparing the average of A and E to the average of J and F: $0.5 * [(a^{++}(\zeta_1) + a^{--}(\zeta_1)) -$

$(a^{++}(\zeta_0) + a^{--}(\zeta_0))]$.

7. Our approach of "controlling for demand effects" can also be extended to weak treatments. In Section 3.4 we outline the conditions under which this approach reduces bias. First, we compare respondents in the treatment and control group who all receive weak positive or weak negative demand treatments. In Figure A.5 the positive-positive point estimate is defined by points I and D: $(a^{+}(\zeta_1) - a^{+}(\zeta_0))$, while the negative-negative estimate is comes from points F and B: $(a^{-}(\zeta_1) - a^{-}(\zeta_0))$.

8. Finally, fully informative demand treatments can be used to eliminate nuisance parameters due to unobservable beliefs, facilitating the estimation of structural models. Structural estimation leverages points A, E, F, and J to estimate model parameters, and uses those to impute the latent demand beliefs at points C and H.

Figure A.5: Using the method in practice: Example



Notes: Figure displays mean actions under different treatment conditions and different demand treatments. Point A is given by respondents in the control group who receive the negative strong demand treatment: $a^{--}(\zeta_0)$; Point B is given by respondents in the control group who receive the negative weak demand treatment: $a^{-}(\zeta_0)$; Point C is defined by respondents in the control group who receive no demand treatment: $a^{L}(\zeta_0)$; Point D is given by respondents in the control group who receive the positive weak demand treatment: $a^{+}(\zeta_0)$; Point E is given by respondents in the control group who receive the positive strong demand treatment: $a^{++}(\zeta_0)$. Points F to J are defined analogously for respondents in the treatment group (ζ_1).

2.C Balance tables, summary statistics, and attrition

Table A.13: Balance Table: Experiment 1 (Strong Demand)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.511	0.519	0.497	0.639	0.442	0.215	4479
Income	51560.364	52459.736	53429.878	0.387	0.068	0.346	3995
Age	36.226	36.464	36.414	0.560	0.653	0.904	4479
Household Size	3.711	3.649	3.626	0.225	0.097	0.640	4479
White	0.773	0.785	0.774	0.429	0.974	0.451	4479
Black	0.070	0.067	0.071	0.676	0.927	0.612	4479
Hispanic	0.053	0.057	0.055	0.592	0.824	0.757	4479
Asian	0.079	0.063	0.075	0.091	0.703	0.194	4479
Full-time employment	0.484	0.508	0.522	0.194	0.039	0.430	4479
Part-time employment	0.128	0.120	0.115	0.541	0.280	0.632	4479
Unemployed	0.144	0.133	0.130	0.409	0.269	0.770	4479
Bachelor Degree	0.353	0.369	0.389	0.360	0.044	0.264	4479
Conservative	0.230	0.238	0.242	0.641	0.459	0.778	4441
Number of HITs	9393.289	9217.178	8651.406	0.762	0.202	0.321	4479
Joint							

Notes: In this table we present evidence on the experimental integrity in experiment 1. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.9110. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.6965. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.2402.

Table A.14: Balance Table: Experiment 2 (Weak Demand)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.466	0.467	0.477	0.949	0.630	0.678	2950
Income	51017.241	51410.405	51949.829	0.757	0.455	0.664	2612
Age	35.909	35.871	35.227	0.940	0.171	0.197	2950
Household Size	3.696	3.686	3.757	0.880	0.346	0.275	2950
White	0.785	0.761	0.749	0.194	0.060	0.568	2950
Black	0.069	0.076	0.076	0.525	0.517	0.994	2950
Hispanic	0.054	0.051	0.057	0.756	0.723	0.507	2950
Asian	0.066	0.070	0.087	0.705	0.083	0.178	2950
Full-time employment	0.493	0.464	0.467	0.199	0.236	0.915	2950
Part-time employment	0.130	0.099	0.125	0.033	0.742	0.071	2950
Unemployed	0.101	0.139	0.129	0.010	0.056	0.493	2950
Bachelor Degree	0.367	0.353	0.376	0.524	0.662	0.283	2950
Conservative	0.273	0.254	0.243	0.342	0.131	0.583	2927
Number of HITs	5854.863	5642.157	5306.841	0.703	0.314	0.529	2950

Notes: In this table we present evidence on the experimental integrity in experiment 2. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.7084. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.2332. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.4838.

Table A.15: Balance Table: Experiment 3 (Effort Experiment with strong demand)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.558	0.574	0.536	0.598	0.439	0.235	1691
Income	33596.974	32213.115	32457.983	0.170	0.265	0.822	1691
Age	37.444	37.359	36.618	0.906	0.251	0.352	1691
Household Size	3.750	3.783	3.771	0.690	0.790	0.894	1691
White	0.752	0.783	0.761	0.217	0.749	0.411	1691
Black	0.109	0.084	0.084	0.149	0.152	0.999	1691
Hispanic	0.055	0.025	0.046	0.006	0.493	0.070	1691
Asian	0.065	0.072	0.074	0.634	0.555	0.914	1691
Full-time employment	0.509	0.498	0.536	0.707	0.364	0.241	1691
Part-time employment	0.125	0.125	0.107	0.993	0.337	0.387	1691
Unemployed	0.105	0.121	0.107	0.380	0.886	0.502	1691
Bachelor Degree	0.395	0.355	0.370	0.155	0.382	0.623	1691
Republican	0.250	0.289	0.273	0.139	0.382	0.585	1691

Notes: In this table we present evidence on the integrity of the randomisation in experiment 3. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.9171. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.1012. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.4845.

Table A.16: Balance Table: Experiment 4 (Representative Sample)

	No demand	Pos. demand	Neg. demand	P-value(Pos. demand - no demand)	P-value(Neg. demand - no demand)	P-value(Pos. demand - neg. demand)	Observations
Male	0.488	0.485	0.468	0.912	0.446	0.428	2933
Income	68357.264	65309.037	67175.470	0.253	0.662	0.398	2882
Age	47.972	46.899	47.879	0.195	0.911	0.147	2933
Household Size	3.332	3.312	3.333	0.752	0.983	0.692	2926
White	0.801	0.772	0.784	0.159	0.399	0.505	2927
Black	0.070	0.069	0.062	0.968	0.518	0.457	2927
Hispanic	0.051	0.064	0.062	0.269	0.379	0.799	2927
Asian	0.043	0.061	0.062	0.104	0.079	0.866	2927
Full-time employment	0.499	0.485	0.496	0.566	0.888	0.604	2933
Part-time employment	0.074	0.078	0.091	0.768	0.218	0.263	2933
Unemployed	0.068	0.050	0.052	0.132	0.188	0.818	2933
Bachelor Degree	0.331	0.352	0.330	0.368	0.975	0.262	2933
Conservative	0.350	0.352	0.351	0.921	0.962	0.951	2797

Notes: In this table we present evidence on the integrity of the randomisation in experiment 4. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.7455. The p-value of the joint F-test when comparing covariates in the positive and no-demand demand condition is 0.4909. The p-value of the joint F-test when comparing covariates in the negative and no-demand demand condition is 0.6390.

Table A.17: Balance Table: Experiment 5 (Many Task experiment)

	Pos. demand	Neg. demand	P-value(Pos. demand - neg. demand)	Observations
Male	0.453	0.472	0.168	5045
Income	53324.385	52746.322	0.460	4478
Age	37.328	37.207	0.714	5045
Household Size	3.711	3.654	0.159	5045
White	0.770	0.776	0.620	5045
Black	0.078	0.072	0.434	5045
Hispanic	0.048	0.048	0.956	5045
Asian	0.075	0.078	0.773	5045
Full-time employment	0.513	0.516	0.819	5045
Part-time employment	0.115	0.113	0.830	5045
Unemployed	0.126	0.140	0.147	5045
Bachelor Degree	0.376	0.372	0.745	5045
Conservative	0.263	0.257	0.636	5019
Number of HITs	9381.041	8544.300	0.055	5045

Notes: In this table we present evidence on the integrity of the randomisation in experiment 5. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.1990.

Table A.18: Balance Table: Experiment 6 (Effort Experiment with weak demand treatments)

	Pos. demand	Neg. demand	P-value(Pos. demand - neg. demand)	Observations
Male	0.547	0.556	0.803	769
Income	32317.708	32532.468	0.861	769
Age	37.339	37.545	0.807	769
Household Size	3.732	3.681	0.626	769
White	0.755	0.730	0.422	769
Black	0.083	0.083	0.991	769
Hispanic	0.055	0.073	0.306	769
Asian	0.081	0.075	0.780	769
Full-time employment	0.552	0.527	0.491	769
Part-time employment	0.128	0.094	0.132	769
Unemployed	0.125	0.122	0.902	769
Bachelor Degree	0.432	0.379	0.134	769
Conservative	0.266	0.325	0.078	764

Notes: In this table we present evidence on the integrity of the randomisation in experiment 6. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.2562.

Table A.19: Balance Table: Experiment 7 (Within Design)

	Pos. demand	Neg. demand	P-value(Pos. demand - neg. demand)	Observations
Male	0.545	0.610	0.038	999
Income	53645.374	54549.763	0.604	876
Age	34.465	34.439	0.970	999
Household Size	3.533	3.540	0.938	999
White	0.732	0.743	0.696	999
Black	0.078	0.078	0.995	999
Hispanic	0.064	0.049	0.300	999
Asian	0.090	0.109	0.317	999
Full-time employment	0.520	0.569	0.118	999
Part-time employment	0.133	0.092	0.043	999
Unemployed	0.137	0.125	0.592	999
Bachelor Degree	0.408	0.386	0.475	999
Conservative	0.241	0.233	0.776	994

Notes: In this table we present evidence on balance for experiment 7. The p-value of the joint F-test when comparing covariates in the positive and negative demand condition is 0.043.

Table A.20: Summary Statistics: Pooled across all experiments

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.50	0.50	0.00	0.00	1.00	18866
Income	52105.70	32773.39	45000.00	5000.00	225000.00	17303
Age	38.25	13.02	35.00	17.00	116.00	18866
Household Size	3.63	1.40	3.00	2.00	13.00	18859
White	0.77	0.42	1.00	0.00	1.00	18860
Black	0.07	0.26	0.00	0.00	1.00	18860
Hispanic	0.05	0.22	0.00	0.00	1.00	18860
Asian	0.07	0.26	0.00	0.00	1.00	18860
Full-time employment	0.50	0.50	1.00	0.00	1.00	18866
Part-time employment	0.11	0.32	0.00	0.00	1.00	18866
Unemployed	0.12	0.32	0.00	0.00	1.00	18866
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	18866
Conservative	0.27	0.44	0.00	0.00	1.00	16942
Number of HITs	8215.31	14921.14	2500.00	750.00	75000.00	12474

Notes: This table summarizes the main covariates of all respondents across all 6 experiments.

Table A.21: Summary Statistics: Experiment 1 (Strong demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.51	0.50	1.00	0.00	1.00	4479
Income	52481.85	26617.99	55000.00	5000.00	100000.00	3995
Age	36.37	11.26	33.00	19.00	88.00	4479
Household Size	3.66	1.40	3.00	2.00	11.00	4479
White	0.78	0.42	1.00	0.00	1.00	4479
Black	0.07	0.25	0.00	0.00	1.00	4479
Hispanic	0.06	0.23	0.00	0.00	1.00	4479
Asian	0.07	0.26	0.00	0.00	1.00	4479
Full-time employment	0.50	0.50	1.00	0.00	1.00	4479
Part-time employment	0.12	0.33	0.00	0.00	1.00	4479
Unemployed	0.14	0.34	0.00	0.00	1.00	4479
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	4479
Conservative	0.24	0.43	0.00	0.00	1.00	4441
Number of HITs	9091.59	15766.32	2500.00	750.00	75000.00	4479

Notes: This table summarizes the main covariates of all respondents in experiment 1.

Table A.22: Summary Statistics: Experiment 2 (Weak demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.47	0.50	0.00	0.00	1.00	2950
Income	51460.57	26145.92	55000.00	5000.00	100000.00	2612
Age	35.67	11.09	33.00	19.00	81.00	2950
Household Size	3.71	1.43	3.00	2.00	13.00	2950
White	0.77	0.42	1.00	0.00	1.00	2950
Black	0.07	0.26	0.00	0.00	1.00	2950
Hispanic	0.05	0.23	0.00	0.00	1.00	2950
Asian	0.07	0.26	0.00	0.00	1.00	2950
Full-time employment	0.47	0.50	0.00	0.00	1.00	2950
Part-time employment	0.12	0.32	0.00	0.00	1.00	2950
Unemployed	0.12	0.33	0.00	0.00	1.00	2950
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	2950
Conservative	0.26	0.44	0.00	0.00	1.00	2927
Number of HITs	5600.34	12081.52	1500.00	750.00	75000.00	2950

Notes: This table summarizes the main covariates of all respondents in experiment 2.

Table A.23: Summary Statistics: Experiment 3 (Effort Experiment: Strong demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.56	0.50	1.00	0.00	1.00	1691
Income	32877.00	17304.76	35000.00	5000.00	85000.00	1691
Age	37.19	12.32	36.00	21.00	70.00	1691
Household Size	3.77	1.39	4.00	2.00	12.00	1691
White	0.76	0.43	1.00	0.00	1.00	1691
Black	0.09	0.29	0.00	0.00	1.00	1691
Hispanic	0.04	0.20	0.00	0.00	1.00	1691
Asian	0.07	0.25	0.00	0.00	1.00	1691
Full-time employment	0.51	0.50	1.00	0.00	1.00	1691
Part-time employment	0.12	0.33	0.00	0.00	1.00	1691
Unemployed	0.11	0.31	0.00	0.00	1.00	1691
Bachelor Degree	0.38	0.48	0.00	0.00	1.00	1691
Republican	0.27	0.44	0.00	0.00	1.00	1691

Notes: This table summarizes the main covariates of all respondents in experiment 3.

Table A.24: Summary Statistics: Experiment 4 (Representative sample)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.48	0.50	0.00	0.00	1.00	2933
Income	66658.57	52862.72	62500.00	7500.00	225000.00	2882
Age	47.50	16.39	47.00	17.00	116.00	2933
Household Size	3.32	1.26	3.00	2.00	13.00	2926
White	0.78	0.41	1.00	0.00	1.00	2927
Black	0.07	0.25	0.00	0.00	1.00	2927
Hispanic	0.06	0.24	0.00	0.00	1.00	2927
Asian	0.06	0.23	0.00	0.00	1.00	2927
Full-time employment	0.49	0.50	0.00	0.00	1.00	2933
Part-time employment	0.08	0.28	0.00	0.00	1.00	2933
Unemployed	0.05	0.23	0.00	0.00	1.00	2933
Bachelor Degree	0.34	0.47	0.00	0.00	1.00	2933
Conservative	0.35	0.48	0.00	0.00	1.00	2797

Notes: This table summarizes the main covariates of all respondents in experiment 4.

Table A.25: Summary Statistics: Experiment 5 (Many task experiment)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.46	0.50	0.00	0.00	1.00	5045
Income	53034.84	26174.26	55000.00	5000.00	100000.00	4478
Age	37.27	11.72	34.00	17.00	88.00	5045
Household Size	3.68	1.44	3.00	2.00	13.00	5045
White	0.77	0.42	1.00	0.00	1.00	5045
Black	0.07	0.26	0.00	0.00	1.00	5045
Hispanic	0.05	0.21	0.00	0.00	1.00	5045
Asian	0.08	0.27	0.00	0.00	1.00	5045
Full-time employment	0.51	0.50	1.00	0.00	1.00	5045
Part-time employment	0.11	0.32	0.00	0.00	1.00	5045
Unemployed	0.13	0.34	0.00	0.00	1.00	5045
Bachelor Degree	0.37	0.48	0.00	0.00	1.00	5045
Conservative	0.26	0.44	0.00	0.00	1.00	5019
Number of HITs	8966.40	15468.91	2500.00	750.00	75000.00	5045

Notes: This table summarizes the main covariates of all respondents in experiment 5.

Table A.26: Summary Statistics: Experiment 6 (Effort Experiment: Weak demand)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.55	0.50	1.00	0.00	1.00	769
Income	32425.23	16975.09	35000.00	5000.00	85000.00	769
Age	37.44	11.73	35.00	21.00	70.00	769
Household Size	3.71	1.46	3.00	2.00	10.00	769
White	0.74	0.44	1.00	0.00	1.00	769
Black	0.08	0.28	0.00	0.00	1.00	769
Hispanic	0.06	0.24	0.00	0.00	1.00	769
Asian	0.08	0.27	0.00	0.00	1.00	769
Full-time employment	0.54	0.50	1.00	0.00	1.00	769
Part-time employment	0.11	0.31	0.00	0.00	1.00	769
Unemployed	0.12	0.33	0.00	0.00	1.00	769
Bachelor Degree	0.41	0.49	0.00	0.00	1.00	769
Conservative	0.30	0.46	0.00	0.00	1.00	764

Notes: This table summarizes the main covariates of all respondents in experiment 6.

Table A.27: Summary Statistics: Experiment 7 (Within Design)

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.58	0.49	1.00	0.00	1.00	999
Income	54081.05	25778.96	55000.00	5000.00	100000.00	876
Age	34.45	10.73	31.00	19.00	83.00	999
Household Size	3.54	1.39	3.00	2.00	13.00	999
White	0.74	0.44	1.00	0.00	1.00	999
Black	0.08	0.27	0.00	0.00	1.00	999
Hispanic	0.06	0.23	0.00	0.00	1.00	999
Asian	0.10	0.30	0.00	0.00	1.00	999
Full-time employment	0.54	0.50	1.00	0.00	1.00	999
Part-time employment	0.11	0.32	0.00	0.00	1.00	999
Unemployed	0.13	0.34	0.00	0.00	1.00	999
Bachelor Degree	0.40	0.49	0.00	0.00	1.00	999
Conservative	0.24	0.43	0.00	0.00	1.00	994

Notes: This table summarizes the main covariates of all respondents in experiment 7.

Table A.28: Attrition overview by task in the strong demand experiments

	Finished: Time	Finished: Risk	Finished: Ambiguity Aversion	Finished: Effort 0 cent bonus	Finished: Effort 1 cent bonus	Finished: Lying	Finished: Dictator Game	Finished: Ult. Game 1	Finished: Ult. Game 2	Finished: Trust Game 1	Finished: Trust Game 2
Panel A: Unconditional Means											
Positive demand	1.000	1.000	0.995	0.972	0.968	0.995	1.000	1.000	0.995	0.990	1.000
	(0.000)	(0.000)	(0.005)	(0.010)	(0.011)	(0.005)	(0.000)	(0.000)	(0.005)	(0.007)	(0.000)
No demand	0.996	1.000		0.941	0.980		1.000				
	(0.004)	(0.000)		(0.015)	(0.009)		(0.000)				
Negative demand	0.992	0.996	1.000	0.984	0.970	1.000	0.996	1.000	0.990	1.000	1.000
	(0.006)	(0.004)	(0.000)	(0.008)	(0.011)	(0.000)	(0.004)	(0.000)	(0.007)	(0.000)	(0.000)
Panel B: Differential attrition											
Positive - Negative	0.008	0.004	-0.005	-0.012	-0.002	-0.005	0.004		0.005	-0.010	0.000
	(0.006)	(0.004)	(0.005)	(0.013)	(0.016)	(0.005)	(0.004)		(0.008)	(0.007)	(0.000)
Positive - Neutral	0.004			0.031	-0.012						
	(0.004)			(0.018)	(0.014)						
Negative - Neutral	-0.004	-0.004		0.043	-0.009		-0.004				
	(0.007)	(0.004)		(0.017)	(0.014)		(0.004)				
Observations	730	729	405	757	734	366	771	409	424	384	371

Notes: In Panel A we present the proportion of respondents who completed the experiment in the positive, negative and no-demand treatment arms respectively. In Panel B we assess whether there was differential attrition across treatment arms by examining differences in completion rates across demand treatment arms.

Table A.29: Attrition overview by task in the weak demand experiments

	Finished: Time	Finished: Risk	Finished: Ambiguity Aversion	Finished: Effort 0 cent bonus	Finished: Effort 1 cent bonus	Finished: Lying	Finished: Dictator Game	Finished: Ult. Game 1	Finished: Ult. Game 2	Finished: Trust Game 1	Finished: Trust Game 2
Panel A: Unconditional Means											
Positive demand	0.995	1.000	0.990	0.955	0.941	0.995	0.996	1.000	1.000	1.000	1.000
	(0.005)	(0.000)	(0.007)	(0.015)	(0.017)	(0.005)	(0.004)	(0.000)	(0.000)	(0.000)	(0.000)
No demand		0.993					0.992				
		(0.005)					(0.006)				
Negative demand	0.991	0.992	0.995	0.965	0.960	1.000	0.993	0.995	1.000	1.000	0.985
	(0.007)	(0.006)	(0.005)	(0.013)	(0.014)	(0.000)	(0.005)	(0.005)	(0.000)	(0.000)	(0.009)
Panel B: Differential attrition											
Positive - Negative	0.004	0.008	-0.005	-0.010	-0.019	-0.005	0.004	0.005		0.000	0.015
	(0.008)	(0.006)	(0.009)	(0.019)	(0.022)	(0.005)	(0.007)	(0.005)		(0.000)	(0.009)
Positive - Neutral		0.007					0.004				
		(0.005)					(0.007)				
Negative - Neutral		-0.001					0.001				
		(0.008)					(0.008)				
Observations	425	743	393	404	401	413	763	361	411	352	349

Notes: In Panel A we present the proportion of respondents who completed the experiment in the positive, negative and no-demand treatment arms respectively. In Panel B we assess whether there was differential attrition across treatment arms by examining differences in completion rates across demand treatment arms.

Chapter 3

Public Debt and the Demand for Government Spending and Taxation

Christopher Roth[1] Johannes Wohlfart

Abstract We examine how beliefs about the debt-to-GDP ratio affect people's attitudes towards government spending and taxation. Using a representative sample of the US population, we provide half of our respondents with information about the debt-to-GDP ratio in the US. We find that most people underestimate the debt-to-GDP ratio and favour a cut in government spending once they learn about the actual amount of debt, but do not alter their attitudes towards taxation. The treatment effects operate through changes in

[1]Christopher Roth: Department of Economics, University of Oxford; Johannes Wohlfart: Goethe University Frankfurt. We would like to thank Jan Bakker, Cornelius Christian, Jon de Quidt, Alexis Grigorieff, Olga Goldfayn, Thomas Graeber, Michalis Haliassos, Johannes Haushofer, Lukas Hensel, Johannes Hermle, Chaning Jang, Yigitcan Karabulut, Hannah Paule-Paludkiewicz, Simon Quinn, Sonja Settele as well as various seminar participants. Financial support from Goethe University Frankfurt is gratefully acknowledged. The experiment is registered in the AEA RCT Registry as trial 1960 available at: `https://www.socialscienceregistry.org/trials/1960`. The experimental instructions are available at: `https://goo.gl/SMYC9f`. Ethics approval was obtained from the University of Oxford. The usual disclaimer applies.

beliefs about fiscal sustainability and persist in a four-week follow-up.

3.1 An experiment on debt perceptions

Government debt in many of the largest economies in the world has increased over the last few decades. For example, the debt-to-GDP ratio in the United States reached a level of 104.81 percent in 2016. High levels of government debt can have important implications for the tax burden of future generations, the sustainability of public finances, and the possibility of a fiscal crisis. Moreover, recent evidence suggests that the build-up of public debt can have negative effects on output growth (Chudik et al., 2016). While the ultimate effects of government debt on the economy are still being debated among economists, far less attention has been devoted to people's beliefs and preferences regarding government debt. Are individuals' estimates of the debt-to-GDP ratio in line with underlying facts? Do voters have a preference for lowering levels of debt? And how do beliefs about the level of government debt affect attitudes towards government spending and taxation? Answering these questions has important implications for understanding voting behavior, patterns of debt accumulation and the optimal design of government policies.

In this paper, we conduct several pre-registered online experiments in the United States in which we measure people's beliefs and preferences regarding government debt. We first elicit people's beliefs about the debt-to-GDP ratio. Then we provide a random subset of our respondents with information about the debt-to-GDP ratio and study how this affects their at-

titudes towards government spending and taxation measured using both self-reports and behavioural measures. We recruit 800 respondents from an online panel that is representative of the US population in terms of age, income, gender and region. Moreover, we recruit 800 individuals from an online labour market, Amazon Mechanical Turk (MTurk), whom we re-survey four weeks after the main experiment.

We start by documenting a series of stylized facts about people's beliefs and preferences regarding government debt. We show that most individuals underestimate the degree of indebtedness of the US government. The median respondent's estimate of the debt-to-GDP ratio is 56.23 percent, almost 50 percentage points below the actual debt-to-GDP ratio (104.81 percent). Moreover, the median respondent thinks that the government should aim to achieve an even lower debt-to-GDP ratio of 25 percent.

We find that individuals who receive information about the true debt-to-GDP ratio become more likely to consider the prevailing level of government debt as too high and become more supportive of cutting the overall amount of debt by approximately 0.26 of a standard deviation. Moreover, people who receive the information also become significantly more likely to prefer government spending cuts in all spending categories except for defense. Our estimated effect sizes for views on government spending are large and correspond to approximately 0.14 of a standard deviation. However, people's views on taxation are not affected by the information treatment. We provide evidence that people's beliefs about debt also affect their political preferences as measured with a behavioural measure. Specifically, respondents provided with the information donate 0.15 of a standard deviation more to a think tank which advocates reducing government spending.

This is a large effect size and corresponds to 54 percent of the gap in donations between Democrats and Republicans. However, we find no evidence that treated respondents change their willingness to sign a petition in favour of a balanced budget rule.

Do treatment effects persist over time? Using data from the four-week follow-up we show that the information about government debt persistently shifts people's views on cutting government debt and total government spending. The follow-up also shows that respondents in the treatment group have significantly lower biases in beliefs about the debt-to-GDP ratio. This suggests that a substantial part of the effects operate through genuine changes in beliefs about the debt-to-GDP ratio rather than through short-lived emotional responses to the treatment.

We also shed light on the mechanisms through which the perceived level of government debt increases people's support for reducing the amount of government debt through spending cuts. We provide evidence that beliefs about public debt causally affect people's perception of the sustainability of public finances, even though we find no strong evidence of changes in expected government spending and taxation. We interpret this as suggestive evidence that people demand immediate spending reductions as a result of a desire to smooth the consumption of public goods over time. We find no evidence that beliefs about government debt causally affect people's trust in the government or their beliefs about rent-seeking and inefficiencies in the public sector, which could similarly lead to a reduction in desired spending levels.

Our paper adds to the literature on the political economy of government debt which tries to understand why governments accumulate high levels of

debt (Alesina and Passalacqua, 2015; Alesina and Tabellini, 1990; Battaglini and Coate, 2008; Cukierman and Meltzer, 1989; Müller et al., 2016; Persson and Svensson, 1989; Song et al., 2012).[2] We add to this literature by providing the first evidence that biased beliefs about the level of government debt can make voters prefer higher levels of government spending than if they were aware of the true level of debt. Biased beliefs could therefore contribute to the accumulation of suboptimal levels of government debt.[3] Our evidence is also related to Lergetporer et al. (2016) and West et al. (2016), who show that informing people about current levels of government spending on different categories sharply reduces people's support for spending increases in these categories.

We also contribute to the literature on the determinants of people's attitudes towards the size of the government and redistributive policies (Alesina and Giuliano, 2011; Alesina and La Ferrara, 2005; Giuliano and Spilimbergo, 2014). Alesina et al. (2018) provide evidence that left-wing respondents increase their support for government redistribution and policies that promote equality of opportunity when exposed to information about low intergenerational mobility, while right-wing respondents do not adjust their attitudes. Kuziemko et al. (2015) show that people's demand for redistribution is fairly inelastic to information about inequality. Our paper extends this literature by providing the first evidence on the role of people's perceptions of government debt in the formation of attitudes towards the size of the government and redistributive policies, such as spending on income support and social insurance programs. Our findings suggest that people become

[2]Battaglini et al. (2016) provide evidence from an abstract laboratory experiment which studies political distortions in the accumulation of public debt.

[3]Moreover, our paper contributes to a small correlational literature examining whether voters punish governments for running budget deficits, which provides indirect evidence on voters' preferences over government debt (Alesina et al., 2012, 1998; Brender and Drazen, 2008; Peltzman, 1992).

less supportive of such policies if they think that levels of government debt are high. Moreover, we add to the literature on attitudes towards the size of the government by introducing two novel behavioural measures: first, we measure people's willingness to donate money to an NGO advocating the downsizing of the government. Second, we capture people's willingness to sign a real online petition in favour of a balanced budget rule.[4]

Finally, by examining whether people's expectations about future taxes and future government spending change in response to changes in beliefs about government debt, we provide evidence on whether our respondents take into account the intertemporal budget constraint of the government when forming their expectations. The assumption that consumers act in line with the intertemporal government budget constraint is at the core of many macro models and is one of the key assumptions underlying the Ricardian Equivalence Theorem (Barro, 1974). We find mixed evidence on whether people form their expectations in line with the intertemporal budget constraint of the government. People who learn about the level of public debt change their beliefs about fiscal sustainability, but do not change their expectations about government spending and taxation in ten years.[5]

The rest of this paper is structured as follows. In section 3.2 we provide some background on the intertemporal budget constraint of the government and the hypotheses we test in the experiment. Section 3.3 presents

[4]More generally, we contribute to a literature on the experimental manipulation of beliefs. For instance, Cruces et al. (2013) and Karadja et al. (2017) provide evidence that people change their demand for redistribution in response to information about their position in the income distribution. Our work is also related to an experimental literature on the formation of expectations about economic variables, such as inflation, house prices and income (Armona et al., 2018).

[5]A small correlational literature has used survey data to examine the assumptions underlying the Ricardian Equivalence Theorem (Allers et al., 1998; Hayo and Neumeier, 2016; Heinemann and Hennighausen, 2012). This literature documents a low level of awareness of the level of public debt and finds little support that individuals' consumption and savings decisions are influenced by the perceived level of debt.

the design as well as the setting and samples used in the information experiments. We first present descriptives in Section 3.4. Then, we show our main results in section 3.5 and provide evidence on mechanisms and robustness in section 3.6. Section 3.7 concludes.

3.2 Conceptual framework

In this section we present a simple conceptual framework which motivates the experiment and the main hypotheses on how voters should adjust their policy preferences when updating their beliefs about the amount of government debt. Voters form their expectations about future government spending and taxation and their policy preferences subject to the perceived intertemporal government budget constraint:

$$\sum_{t=1}^{\infty} \frac{p_t}{(1+r)^t} = \alpha \left[B_0 + \sum_{t=1}^{\infty} \frac{T_t}{(1+r)^t} \right]$$

where p_t is public good provision in period t, T_t is total tax revenue collected in period t and B_0 is net wealth of the government in period 0.[6] α lies in the interval $[0, 1]$ and captures the efficiency of the bureaucratic process. In our experiment we create exogenous variation in our respondents' perceived level of government debt, $-B_0$. If respondents understand the intertemporal budget constraint of the government, then an increase in the perceived level of government debt, $\Delta B_0 < 0$, should lead to a decrease in the perceived net present value of the stream of public good provision, $\sum_{t=1}^{\infty} \frac{p_t}{(1+r)^t}$, or

[6] For simplicity we abstract from the distributional aspects of taxation and from how the tax burden is spread across the electorate.

to an increase in the perceived net present value of tax revenue, $\sum_{t=1}^{\infty} \frac{T_t}{(1+r)^t}$.

If respondents expect that adjustments in spending or taxation will be necessary during their lifetimes, or if they care about the utility of future generations, then an inclination to smooth the consumption of public goods over time could lead them to immediately demand lower levels of government spending in response to learning that debt is higher then they previously thought. Alternatively, respondents could favour immediate tax increases in order to smooth the tax burden for themselves and their children. If respondents do not expect that adjustments will be necessary during their lifetimes, and if they do not care about future generations, they may not respond to the treatment and may want to postpone the necessary adjustments in spending or taxes.

In addition, our respondents could adjust their beliefs about the wastage that occurs in the bureaucratic process. Specifically, upon learning that government debt is higher than they previously thought, respondents could hold more pessimistic beliefs about the rate at which the government transforms tax revenue into public goods, α. Consequently, respondents may want to shift consumption away from public goods towards private goods, and therefore reduce the size of the government.

In section 3.A in the Appendix we demonstrate how an increase in the perceived level of government debt leads to a lower demand for government spending and to an increase in the net present value of total tax revenue collected in a simple two-period model.

3.3 Experimental design

In this section we present our experimental design and explain the structure of the main experiment and the follow-up survey. The full experimental instructions are available at `https://goo.gl/SMYC9f`. We pre-registered our experimental design, the sample sizes, as well as the specifications estimated in the paper on the AEA RCT registry.[7]

3.3.1 Main experiment

Belief elicitation Our experiment is structured as follows: First, we ask all respondents questions about some of their demographics, namely gender, age, region and income, and about their political affiliation. Second, we elicit all participants' beliefs about the debt-to-GDP ratio. In order to make this statistic more meaningful to our respondents, we tell them that government debt refers to the "total amount owed by the Federal government" and that GDP refers to the "market value of all final goods and services that are produced by an economy within one year". Moreover, Ansolabehere et al. (2013) argue that providing people with meaningful anchors increases the data quality and reduces noise in the belief elicitation of abstract statistics. In order to reduce the likelihood that people give random responses as they may not have a sense of the likely magnitude of the debt-to-GDP ratio, we therefore inform our respondents about the debt-to-GDP ratio in 1970 (34.78 percent) before asking them to estimate the current debt-to-GDP ratio in the

[7]The pre-analysis plans are available at `https://www.socialscienceregistry.org/trials/1960`. In this trial entry we also pre-specified an experiment examining how people's support for government spending programs varies by the proposed mode of financing, which will be used as the basis for a separate paper and which is not included in this paper for brevity's sake. Results are available upon request.

US. Our respondents receive the following instructions:

> In 1970 the debt-to-GDP ratio was 34.78 percent. This means that the Federal Government owed around a third of what the country produced within one year. What do you think was the debt-to-GDP ratio in 2016?

We ran a pilot experiment on Amazon Mechanical Turk with 200 participants in order to test how the anchor affects people's beliefs about the level of public debt in 2016. We describe these results in Section 3.4.1.

Information treatment Thereafter, respondents in the treatment group receive information about the actual debt-to-GDP ratio in the US in 2016 (104.81 percent), while respondents in the control group do not receive any information. Treated respondents receive the following message:

> We now would like to provide you with information about the debt-to-GDP ratio in the US. In 2016, the federal debt-to-GDP ratio was 104.81 percent. This means that the Federal Government owed a bit more than what the country produced within one year.

Our respondents also are shown a figure contrasting the debt-to-GDP ratio in 1970 with the debt-to-GDP ratio in 2016 (see Figure A.1). To ensure a high level of trust of our respondents in this statistic we provide them with the official source of the data (the Federal Reserve Bank of St. Louis).

Measuring political preferences: Survey measures To analyze the effects of our information treatment, we first measure our respondents' subjective beliefs about government debt. We ask all of our respondents whether they think that there is too much government debt in the US and whether the government should reduce the amount of debt. We measure people's agreement to these statements on 5-point Likert scales reaching from "Strongly Agree" to "Strongly Disagree". Thereafter, we ask them a series of questions on their attitudes towards government spending. They first answer a question on whether they would like the overall level of government spending to be increased or decreased. Then we provide our respondents with explanations of several spending categories. For each category, we ask them whether they would like to increase or decrease spending.[8] The answer categories for these questions range from 1 "It should be increased a lot" to 5 "It should be decreased a lot". Subsequently, our respondents answer a series of questions on whether income taxes of different income groups should be increased or decreased, whether the government should introduce a wealth tax and whether the estate tax should be increased or decreased.

Measuring political preferences: behavioural measures To examine whether the information also affects political behavior, we employ revealed preference measures of political attitudes. Specifically, we use two (randomly ordered) behavioural measures that capture people's attitudes towards government spending and the size of the government. We employ these measures with our sample from Amazon Mechanical Turk.[9]

[8] We focus on the following spending categories: defense, infrastructure, schooling, social security, social insurance, health, and environment.

[9] We could not include these behavioural measures in the experiment with the representative online panel due to constraints from our online panel provider.

Our respondents can make a donation to an NGO which advocates downsizing the government. Our respondents are told that one out of 20 participants will be chosen to receive an additional $5 at the end of the experiment, and they have to decide how much to keep for themselves and how much to donate. We believe that this is a particularly suitable behavioural measure as donations to political organisations are an important real-life tool for people to support particular political causes.

Moreover, we give all of our respondents the opportunity to sign a real online petition on the White House webpage in favour of introducing a balanced-budget rule. Individuals in the treatment and in the control group receive different links to identical petitions. This allows us to observe the actual numbers of signatures for the petition for people in the treatment group and for people in the control group.

Mechanisms: Post-treatment beliefs To understand why our respondents may change their views on government spending and taxation, we collect a rich set of post-treatment beliefs. Specifically, we measure our respondents' expectations about future taxation and government spending as well as their beliefs about the sustainability of public finances. For example, we ask them whether they agree that "the current levels of government spending and taxation are not sustainable". Thereafter, we measure our respondents' trust towards the government and their beliefs about the effectiveness of the government and about the corruption of politicians. For example, we ask them whether they agree that "the government makes good use of taxpayers' money". Finally, we ask our respondents a series of questions on their demographics.

3.3.2 Follow-up survey

One concern with the evidence from the main survey could be that responses to the information treatment are very short-lived in nature and do not persist over time. To examine the persistence of effects over time, we conduct a follow-up survey four weeks after the main experiment, in which we do not administer any additional treatment. We ask our respondents the same set of questions on their views regarding government spending and taxation. At the very end of the follow-up survey we also ask people about their estimate of the current debt-to-GDP ratio to test whether posterior beliefs about the debt-to-GDP ratio persistently respond to the information.

3.3.3 Setting and sample size

Sample: Representative online panel Our main evidence comes from an experiment with a representative online sample that we conducted in February 2017. We collect data through an online survey in collaboration with the market research company Research Now which is widely used in economics research. This sample of 813 respondents is representative of the adult US population in terms of region, income, age, and gender. Table A.4 displays summary statistics for our sample and the American Community Survey. 55 percent of our respondents are female, a slightly larger fraction than among the American population (51 percent). Moreover, our sample is very similar to the US population in terms of the age profile and regions. While the mean household income in our sample ($62,487) is lower than that of the US population ($84,568), the median household income in our sample ($62,500) is very close to the median household income in the US ($59,039). The attri-

tion rate in our experiment is very low and does not differ across treatment arms.

Sample: Amazon Mechanical Turk In addition to conducting our experiment with the representative online panel, we also recruit participants from Amazon Mechanical Turk (MTurk), an online labour market which is increasingly used to conduct experimental research (Cavallo et al., 2016; Kuziemko et al., 2015). We conduct our experiment on the MTurk platform for several reasons: first, it allows us to employ behavioural measures which are difficult to implement with representative online panels. Second, it enables us to conduct a four-week follow-up with a much higher response rate compared to the response rates of follow-up surveys from the representative online panel.

We ran the experiment on Amazon Mechanical Turk (MTurk) on the 27th of January 2017. We recruit participants currently living in the United States who have completed at least 500 tasks with an overall rating of more than 95 percent. In our experiment on MTurk we recruited 802 participants who completed the experiment. The attrition rate is below 2 percent and not statistically different for respondents in the treatment and the control group. Table A.2 in the Appendix summarizes the characteristics of the sample. 56 percent of our respondents are male, the median income in our sample is $62,500 which is only slightly higher than the median income in the US population. Respondents from the MTurk sample are more educated than respondents from the general population and there are more Democrats in our MTurk sample compared to the US population. 74 percent of the respondents who completed the main experiment also completed our four-week follow-up survey. The sample composition is virtually unchanged

compared to the main experiment and the attrition rate from the main experiment to the follow-up is similar and not statistically different between treatment and control group.

Integrity of randomisation We provide evidence that our representative online panel is balanced between treatment and control group in terms of observables (see Table A.5). We find balance for all covariates except for the fraction of Republicans which is significantly higher in the control group. A joint F-test when regressing the treatment indicator on all covariates confirms that covariates are globally balanced (p=0.243). We deal with the imbalance of the Republican indicator by including all covariates as control variables in our main specification. Note that the larger fraction of Republicans in the control group works against finding a significant effect of the information on views on government spending, given that Republicans are generally more likely to demand spending cuts.

Moreover, the MTurk sample is balanced in terms of the pre-specified observables (see Table A.6). Indeed, we find no imbalance for any of the pre-specified covariates included in the balance test. We also find that the covariates are not jointly significantly related to the treatment indicator (p=0.1311). Finally, we find balance also for respondents in our follow-up survey (see Table A.7). Specifically, we find that all 13 variables in our balance test except one (full-time employment) are balanced. A joint F-test when regressing the treatment indicator on all covariates confirms that covariates are globally balanced (p=0.333).

3.4 Beliefs about the debt-to-GDP ratio

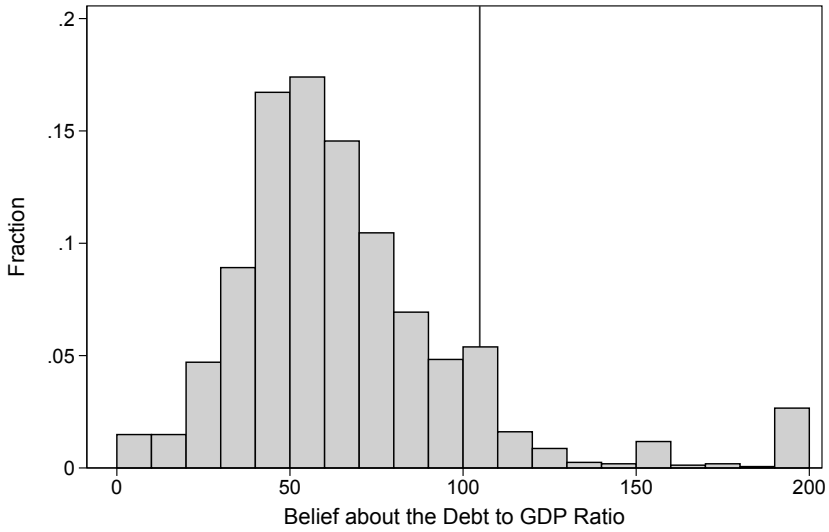
3.4.1 Prior beliefs

The US debt-to-GDP substantially increased over the last decades from about 35 percent in the 1960s and 1970s to more than 100 percent today (see Figure A.2 in the Appendix). Figure 3.1 shows data characterizing people's beliefs about the debt-to-GDP ratio in 2016. We find that people widely underestimate the actual debt-to-GDP ratio in 2016 (104.81 percent). The median respondent believes that the debt-to-GDP ratio is 55.10 percent and more than 90 percent of our respondents underestimate the debt-to-GDP ratio.[10] These findings are consistent with previous evidence that voters have incorrect perceptions of the level of government debt (Allers et al., 1998; Mayer, 1995).

We provide respondents with a historical anchor in order to make sure that their estimate of the debt-to-GDP ratio today and the information treatment are meaningful to them. However, it could be the case that our respondents underestimate the debt-to-GDP ratio because they are given the historical anchor. To rule out this concern we ran a pilot experiment on MTurk with 200 respondents in which we elicited people's beliefs about the debt-to-GDP ratio. Half of the respondents are given information about the debt-to-GDP ratio in 1970, while the remaining respondents do not receive this information. The patterns are very similar for respondents that do not receive the historical anchor. While the median respondent thinks that the debt-to-GDP ratio is 61.5 percent when not provided with an anchor, the median respon-

[10] As pre-specified, we winsorize people's beliefs about the debt-to-GDP ratio at 200 in order to deal with outliers.

Figure 3.1: Beliefs about the debt-to-GDP ratio (pooled sample)



Notes: In this figure we display people's beliefs about the current debt-to-GDP ratio using data on 1612 respondents from the representative online panel and MTurk. The estimates are winsorised at a debt-to-GDP ratio of 200 percent.

dent who is given the historical anchor believes that the true value is 56.23 percent. Even though this difference is statistically significant, it is of small magnitude. Overall, the provision of the anchor reduces the dispersion of beliefs and leads to a lower number of outliers (see Figure A.5).

3.4.2 What is our respondents' desired debt-to-GDP ratio?

In the pilot experiment on MTurk we also ask people about their views on what debt-to-GDP ratio the government should aim to achieve. People answer this question after estimating the current debt-to-GDP ratio in the United States. Figure A.4 shows the distribution of desired changes in the debt-to-GDP ratio, which is defined as the difference between people's desired debt-to-GDP ratio and their belief about the actual debt-to-GDP ratio. The figure highlights that people predominantly want to reduce the amount

of debt in the US. Figure A.3 displays the distribution of beliefs about the debt-to-GDP ratio as well as the desired debt-to-GDP ratio in the group of respondents who received the historical anchor. While the median respondent's estimate of the debt-to-GDP ratio is 56.23 percent, she thinks that the government should aim to achieve a debt-to-GDP ratio of 25 percent.[11]

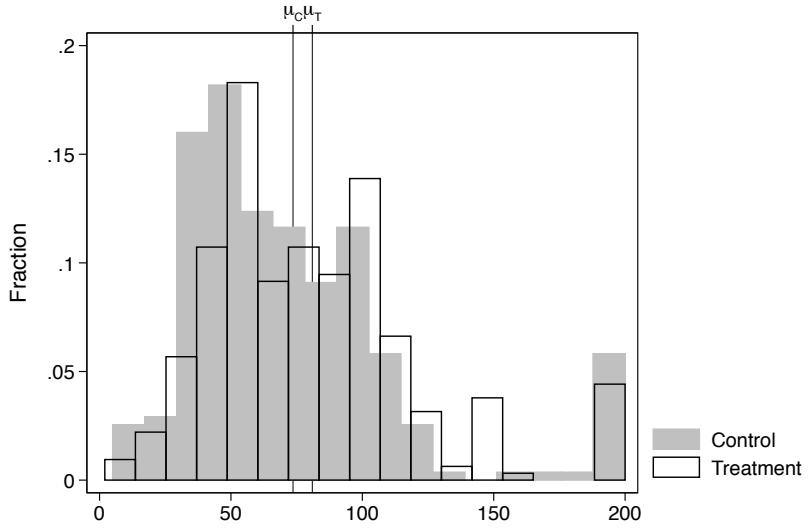
3.4.3 Do respondents update their beliefs?

Do our respondents persistently update their beliefs about the debt-to-GDP ratio in response to the provision of official statistics? We provide evidence that our information treatment durably shifts people's beliefs about the debt-to-GDP ratio using data from the four-week follow-up survey on MTurk. The distribution of posterior beliefs is described in Figures 3.2 and A.7. People in the treatment group report significantly higher estimates of the debt-to-GDP ratio (p=0.002). The median belief in the treatment group is that the debt-to-GDP ratio is 75 percent, while it is 62 percent in the control group. Figure 3.3 shows treatment effects on posterior beliefs depending on our respondents' prior beliefs. The figure highlights that treated subjects who under-estimated the debt-to-GDP ratio strongly shift their belief upward, while treated respondents who over-estimated the debt-to-GDP ratio shift their belief downward (although this effect is noisily measured). This evidence strongly suggests that the information treatment leads to genuine updating of beliefs.

Moreover, Figure A.8 depicts updating of beliefs about the debt-to-GDP ratio using the posterior belief in the follow-up survey and the prior be-

[11] People's stated preference over their desired debt-to-GDP ratio could be much higher if people fully understood what moving to a lower debt-to-GDP ratio would entail in terms of spending cuts or tax increases.

Figure 3.2: Beliefs about the debt-to-GDP ratio in the four-week follow-up

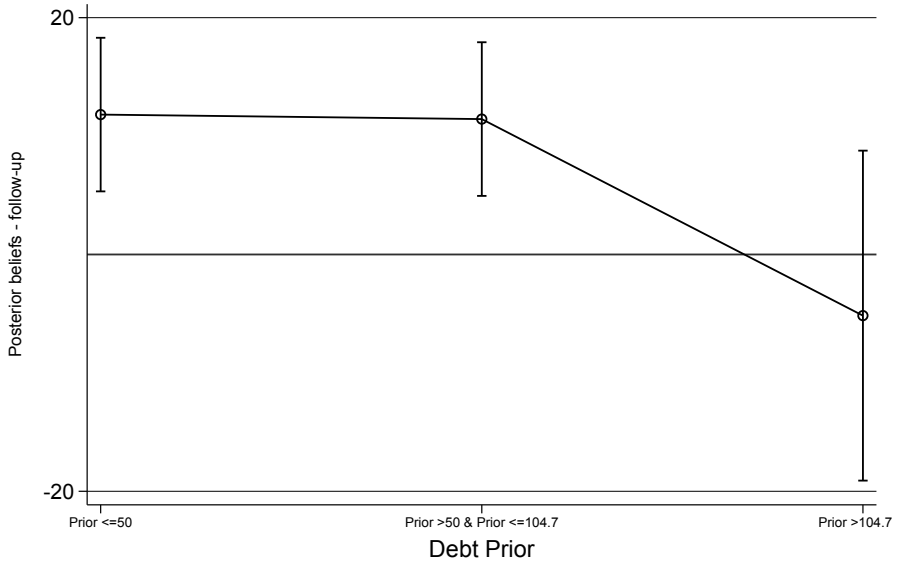


Notes: This figure describes the distribution of beliefs about the debt-to-GDP ratio in the four-week follow-up experiment for the treatment and the control group. This is based on 592 respondents who completed the follow-up. The estimates are winsorised at a debt-to-GDP ratio of 200 percent. The median belief in the treatment group is that the debt-to-GDP ratio is 75 percent, while it is 62 percent in the control group. A Kolmogorov–Smirnov test reveals that the distribution of beliefs is statistically different between the treatment and control group (p=0.018). Also the mean belief about the debt-to-GDP ratio in the treatment group is statistically different from the mean in the control group (p=0.001).

lief from the main survey. We observe some changes in beliefs also in the control group. However, we cannot reject the equality of means of control group beliefs about debt-to-GDP in the main survey and the follow-up survey (p=0.107).

We also examine which covariates predict persistent shifts in beliefs in response to our treatment. We regress the change in the estimated debt-to-GDP ratio between the initial survey and the follow-up survey in the treatment group on a set of demographic variables and people's prior estimate of the debt-to-GDP ratio. Table A.8 shows that people with a higher prior belief about the debt-to-GDP ratio update less, and that male respondents update more. However, none of the other covariates, such as people's polit-

Figure 3.3: Heterogeneous effects on posterior beliefs about the debt-to-GDP ratio: by prior beliefs



Notes: This figure describes treatment effects on posterior beliefs about the debt-to-GDP ratio by people's prior beliefs about the debt-to-GDP ratio. The figure displays the point estimate of the treatment effects with 90 percent confidence interval estimated on data from the follow-up survey on MTurk. The treatment effect estimates control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, and the respondent's number of children (top-coded at five).

ical affiliation, is significantly correlated with the degree of belief updating. Taken together, we find that our treatment leads to strong and persistent updating of beliefs about the debt-to-GDP ratio towards the true value, and the change in beliefs is greater for individuals with more biased beliefs before the treatment.

3.4.4 Correlates of beliefs about the debt-to-GDP ratio

To shed light on the determinants of respondents' beliefs about the level of debt, we regress people's perceived debt-to-GDP ratio on a set of demographics.[12] As more than 90 percent of our respondents underestimate the debt-to-GDP ratio, higher estimates largely correspond to less biased beliefs. Men and older individuals report higher estimates of the debt-to-GDP ratio, even though these effects are significant only in the representative sample and only in the MTurk sample, respectively (see Table A.28 in the Appendix). Education and income, by contrast, are not correlated with people's perceived level of debt.

3.5 The causal effect of information about debt

In this Section we describe the results from our information experiments which allow us to provide evidence on the causal effect of beliefs about the debt-to-GDP ratio on people's views on government debt, public spending and taxation.

3.5.1 Empirical specification

We regress our outcome variables y_i on a treatment indicator, Treatment$_i$, which takes the value one for people who receive the information treatment,

[12] We winsorize the estimates of the debt-to-GDP ratio at 200 percent in order to take care of outliers.

and zero otherwise. We estimate the following equation using OLS:

$$y_i = \pi_0 + \pi_1 \text{Treatment}_i + \Pi^T \mathbf{X_i} + \varepsilon_i$$

where X_i is a vector of control variables, including all of the variables we use in the baseline balance check[13] and ε_i is an individual-specific error term. We include control variables as this increases our power to precisely estimate treatment effects and to account for the small imbalance we observe in the representative online panel for the Republican dummy. We report robust standard errors for all estimations.

We report results for all pre-specified outcome variables which are normalised using the mean and standard deviation from the control group. To deal with the issue of multiple hypotheses testing we create indices of outcomes for views on government debt, for government spending and for taxation, respectively, as described in Anderson (2008). Moreover, within each of these families of outcomes we provide p-values that control for a false discovery rate of 5 percent, i.e. the expected proportion of rejections that are type I errors (Anderson, 2008).

[13]Specifically, we control for the belief about the debt-to-GDP ratio pre-treatment, gender, age, log income, the number of children, dummies for employment status, whether the respondent has a college degree and whether the respondent is a Republican. For ease of interpretation and to take care of outliers we deviate in some minor ways from the pre-specified set of controls. Namely, we include a dummy for "other employment status" and we top-code the number of children at five. We also include a dummy indicating missing values for income, a measure of trust in statistics and a dummy variable for Independents. The two latter control variables help us to increase efficiency, while not affecting the coefficient estimates. Results without controls are presented in the Appendix.

3.5.2 Does the information affect views on government debt?

Do people's subjective views on government debt respond to factual information about the debt-to-GDP ratio? Table 3.1 shows that people who received information about the true debt-to-GDP ratio are significantly more likely to think that there is too much government debt and that the government should reduce the overall amount of government debt. The estimated effects are large in magnitude: People become 0.32 of a standard deviation more likely to think that there is too much debt and 0.26 of a standard deviation more inclined to think that the government should reduce the amount of government debt. Moreover, the effects are very similar for the representative online panel and the MTurk sample, which highlights the robustness of our results.

3.5.3 Does the information affect policy views?

After establishing that people who receive the information want to reduce government debt, we now turn to the question whether people would like to achieve the reduction in debt through spending cuts or through tax increases. Table 3.2 highlights that participants who were provided with the information are 0.18 of a standard deviation more in favour of reducing the overall amount of government spending. People become significantly more likely to demand spending cuts on infrastructure, schooling, social services, health and the environment, but not on defense. Overall, we find fairly large effect sizes of about 0.14 of a standard deviation. Our treatment shifts policy preferences by one third of the preference gap for these variables between Republicans and Democrats. Even though the effect sizes are larger for the

representative online panel than for the MTurk sample, these differences are not statistically distinguishable.

Moreover, we examine people's views on taxation. Table 3.3 shows that people who learn about the true debt-to-GDP ratio become marginally significantly more likely to favour an increase in the overall amount of taxes collected by the government. While treated respondents in the representative sample favour an increase in the estate tax, treated respondents in the MTurk sample become more supportive of increasing income taxes for the bottom 50 percent. However, these results are not robust to adjusting p-values for multiple hypothesis testing. All in all, our results suggest that learning that the debt-to-GDP ratio is higher than previously thought makes people more supportive of government spending cuts, but does not strongly change their support for changes in taxation. The differential responses for government spending and taxation could be due to several factors. First, the perceived marginal disutility of a tax increase could be higher than the perceived marginal disutility of a government spending cut. For instance, this could be due to people's belief that a large fraction of government spending is wasteful. Second, tax increases affect some people's income with certainty, while it is less clear whether individuals will be directly affected by cuts in government spending.[14]

[14]In unreported regressions we found no evidence that treated respondents favour tax increases in other income groups than their own (results available upon request).

Table 3.1: Views on government debt

	There is too much debt	Gov. should reduce debt	Index
Panel A: Pooled			
Treatment	0.326***	0.259***	0.280***
	(0.044)	(0.045)	(0.043)
Adjusted p-value	[0.001]	[0.001]	
Observations	1614	1614	599
Panel B: Rep. Sample			
Treatment	0.346***	0.293***	0.309***
	(0.063)	(0.062)	(0.060)
Adjusted p-value	[0.001]	[0.001]	
Observations	812	812	812
Panel C: MTurk			
Treatment	0.301***	0.218***	0.245***
	(0.063)	(0.065)	(0.062)
Adjusted p-value	[0.001]	[0.001]	
Observations	802	802	802
Panel D: Follow-up Sample			
Treatment	0.159**	0.177**	0.168**
	(0.075)	(0.073)	(0.070)
Adjusted p-value	[0.033]	[0.033]	
Observations	599	599	599

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. In Panel D we show the results from the MTurk follow-up. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table 3.2: Attitudes towards government spending

	Reduce Overall Sp.	Reduce Defense Sp.	Reduce Infrastr. Sp.	Reduce Schooling Sp.	Reduce Social Sec. Sp.	Reduce Social Ins. Sp.	Reduce Health Sp.	Reduce Environm. Sp.	Index
Panel A: Pooled									
Treatment	0.177***	0.011	0.142***	0.186***	0.174***	0.152***	0.156***	0.137***	0.142***
	(0.046)	(0.046)	(0.049)	(0.046)	(0.046)	(0.045)	(0.046)	(0.044)	(0.026)
Adjusted p-value	[0.001]	[0.111]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	
Observations	1613	1613	1613	1613	1613	1613	1613	1613	1613
Panel B: Rep. Sample									
Treatment	0.184***	0.045	0.217***	0.219***	0.212***	0.137**	0.191***	0.164**	0.171***
	(0.068)	(0.066)	(0.070)	(0.069)	(0.068)	(0.066)	(0.067)	(0.065)	(0.039)
Adjusted p-value	[0.007]	[0.065]	[0.005]	[0.005]	[0.005]	[0.015]	[0.005]	[0.009]	
Observations	811	811	811	811	811	811	811	811	811
Panel C: MTurk									
Treatment	0.163**	-0.004	0.052	0.131**	0.100	0.153**	0.094	0.093	0.098***
	(0.063)	(0.063)	(0.068)	(0.064)	(0.063)	(0.063)	(0.063)	(0.060)	(0.036)
Adjusted p-value	[0.067]	[0.313]	[0.222]	[0.086]	[0.128]	[0.067]	[0.128]	[0.128]	
Observations	802	802	802	802	802	802	802	802	802
Panel D: Follow-up Sample									
Treatment	0.156**	-0.025	0.047	0.025	0.077	0.085	0.073	-0.042	0.050
	(0.073)	(0.074)	(0.082)	(0.074)	(0.079)	(0.074)	(0.071)	(0.070)	(0.043)
Adjusted p-value	[0.339]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	
Observations	599	597	597	597	597	597	597	597	597

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. In Panel D we show the results from the MTurk follow-up. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table 3.3: Attitudes towards taxation

	Increase overall amount of taxes	Increase income tax: top 10	Increase income tax: next 40	Increase income tax: bottom 50	Introduce wealth tax	Increase estate tax	Index
Panel A: Pooled							
Treatment	0.091**	0.015	0.002	0.049	-0.040	0.063	0.024
	(0.046)	(0.047)	(0.048)	(0.049)	(0.048)	(0.047)	(0.028)
Adjusted p-value	[0.421]	[1.000]	[1.000]	[0.999]	[0.999]	[0.807]	
Observations	1613	1612	1612	1612	1611	1610	1610
Panel B: Rep. Sample							
Treatment	0.061	-0.027	-0.045	-0.036	0.037	0.134**	0.010
	(0.068)	(0.069)	(0.069)	(0.072)	(0.066)	(0.068)	(0.040)
Adjusted p-value	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[0.399]	
Observations	811	810	810	810	809	809	809
Panel C: MTurk							
Treatment	0.131**	0.067	0.050	0.135**	-0.096	-0.003	0.047
	(0.062)	(0.064)	(0.066)	(0.068)	(0.069)	(0.067)	(0.039)
Adjusted p-value	[0.171]	[0.422]	[0.489]	[0.171]	[0.280]	[0.803]	
Observations	802	802	802	802	802	801	801
Panel D: Follow-up Sample							
Treatment	0.030	-0.138*	0.012	0.151*	0.144*	-0.085	0.017
	(0.074)	(0.082)	(0.075)	(0.079)	(0.080)	(0.079)	(0.039)
Adjusted p-value	[0.691]	[0.222]	[0.733]	[0.222]	[0.222]	[0.268]	
Observations	597	597	597	597	596	593	593

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. In Panel D we show the results from the MTurk follow-up. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

3.5.4 Does the information affect behavior?

To examine whether the information also changes actual political behavior, we analyze our respondents' inclination to donate to a political NGO advocating government spending cuts and their willingness to sign real online petitions. We find that respondents who receive the information donate significantly more money to an NGO lobbying for downsizing the government in the United States. Table 3.4 shows that donations increase by 0.15 of a standard deviation, which corresponds to 54 percent of the gap in donations between Republicans and Democrats. People in the control group donate on average around 58 cents of the 5 dollars, while people in the treatment group donate around 72 cents on average, i.e. donations increase by 24 percent.

However, treated respondents do not become more willing to sign a petition in favour of introducing a balanced budget rule. Table 3.4 highlights that the effects are of small size and statistically insignificant for the self-reported intention to sign the petition. The actual signatures on the petition websites for the treatment and the control group confirm the results from the self-reports.

3.5.5 Do the treatment effects persist?

One concern with survey experiments is that treatment effects could reflect short-lived emotional responses to the information or experimenter demand rather than true changes in beliefs and policy views. Following Cavallo et al. (2016), we address these concerns by examining the persistence of our main

Table 3.4: behavioural measures

	Petition in favor of a balanced budget rule				Donation to
	Want to sign	Report: Signed	Index: Self-reports	Actual signatures	Cato Institute
	(1)	(2)	(3)	(4)	(5)
Treatment	0.0419	-0.0105	0.00659	0.031	0.154**
	(0.032)	(0.028)	(0.027)	(0.019)	(0.077)
Control group mean	0.000	0.000	0.000	0.060	0.000
Observations	802	802	802	802	802

Notes: The outcome variables in columns 1, 2, 3 and 5 are z-scored using the mean and standard deviation in the control group. All estimations are based on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). Robust standard errors are in parentheses.

results in the MTurk sample in a four-week follow-up.[15] We first show that the effects on views regarding government debt persist and remain very large in magnitude. As shown in Panel D of Table 3.1, even four weeks after receiving the treatment respondents remain 0.16 of a standard deviation more likely to think that there is too much debt and 0.18 of a standard deviation more inclined to think that the government should reduce the amount of government debt.[16] This corresponds to approximately 60 percent of the effect size in the main study. Moreover, we find a persistent treatment effect of 0.15 of a standard deviation on people's attitudes towards cutting the overall amount of government spending (Panel D of Table 3.2). This effect size corresponds to 83 percent of the effect size in the main study.

Even though the effects become weaker and are not significantly different from zero for the individual spending categories, they are statistically indistinguishable from the effects in the main experiment. It is worth noting

[15]Haaland and Roth (2017) employ 'demand treatments' (de Quidt et al., 2018) to provide evidence that follow-up surveys are an effective way of dealing with experimenter demand effects.

[16]In Tables A.13 - A.15 in the Appendix we provide evidence on the persistence of results controlling for sample composition effects.

that the effect sizes on individual spending categories estimated in the main study were slightly smaller in the MTurk sample than in the representative sample to begin with. Moreover, Table 3.3 shows little persistence of the effects on people's views on whether to increase the overall amount of taxes. We find suggestive evidence that treated respondents favour an introduction of a wealth tax and an increase in taxes for the bottom 50 percent when re-interviewed after four weeks. However, none of these results are robust to adjusting p-values for multiple hypothesis testing. Naturally, our evidence from the follow-up is not as highly powered as the evidence from the main experiment as we only conducted the follow-up on MTurk where we successfully recontacted 75 percent of the original sample. Taken together, the fact that most of our findings persist in a four-week follow-up suggests that our results reflect true updating of beliefs and policy views, and that short-lived responses to our treatment are not an important concern.

3.5.6 Is there a heterogeneous response to the information?

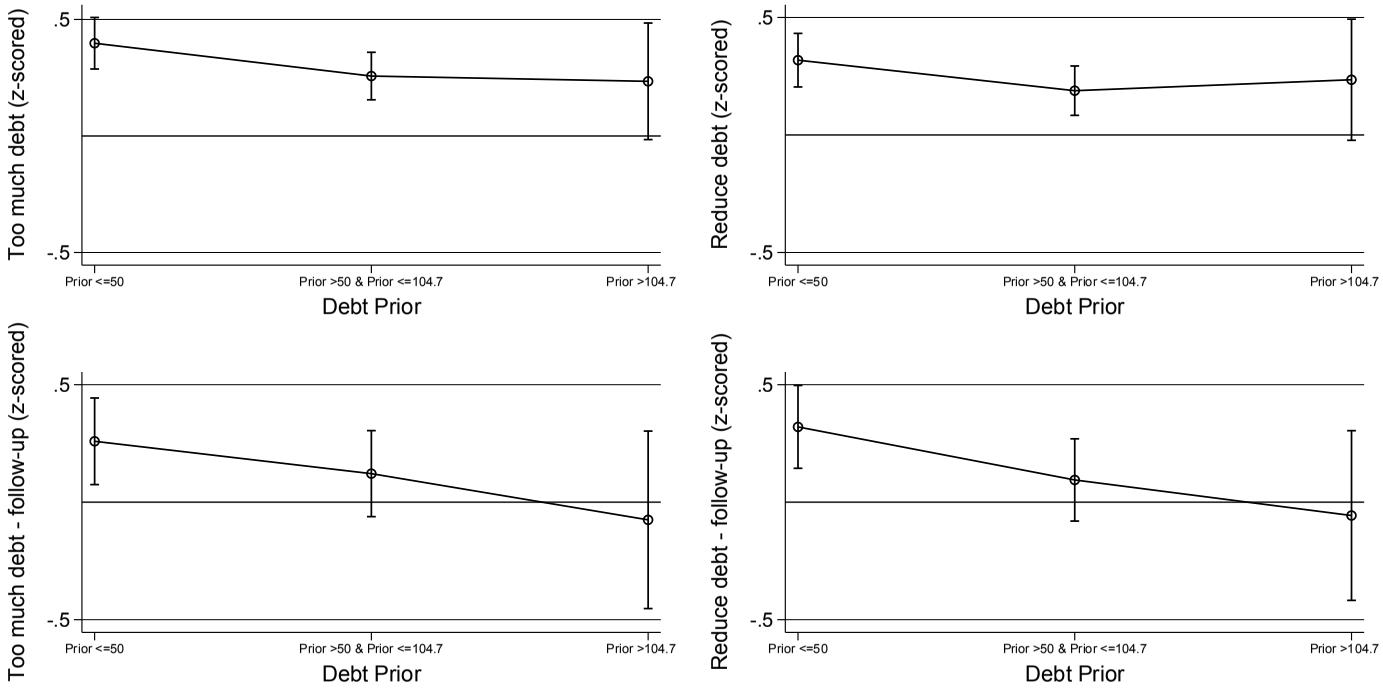
Our information treatment is designed to be more effective for people who have highly biased beliefs about the debt-to-GDP ratio. Respondents who reported lower estimates of the debt-to-GDP ratio ex-ante respond more strongly to our information treatment. In Figures 3.4, 3.5 and A.9 we examine the treatment effects by prior beliefs. Respondents with prior beliefs of a debt-to-GDP ratio below 50 percent respond most strongly to the information in terms of their views on debt reduction and government spending. For respondents who initially over-estimated the debt-to-GDP ratio and receive the treatment, on the other hand, we find noisily measured null effects. The fact that the treatment effects are stronger for individuals with a lower

prior belief and absent for individuals with unbiased prior beliefs suggests that our results reflect true updating of beliefs rather than emotional responses and priming effects.

In Table A.16 in the Appendix we also present heterogeneous treatment effects according to a winsorised measure of people's prior belief about the debt-to-GDP ratio.[17] We include an interaction term of this measure with the treatment dummy and also add it as a control. While the estimated coefficients on the interaction term go into the expected direction, they are not statistically significant. We believe that there are two main reasons for the lack of significance. First, the bias is noisily measured as there is substantial variation in people's ability to estimate abstract statistics. Second, the size of the bias is correlated with many unobserved variables which could affect the response to the information treatment. For example, one could imagine that more highly biased respondents are also less numerate and thus less capable of using our information to update their beliefs. This could lead to a downward bias of the estimated coefficient on the interaction term.

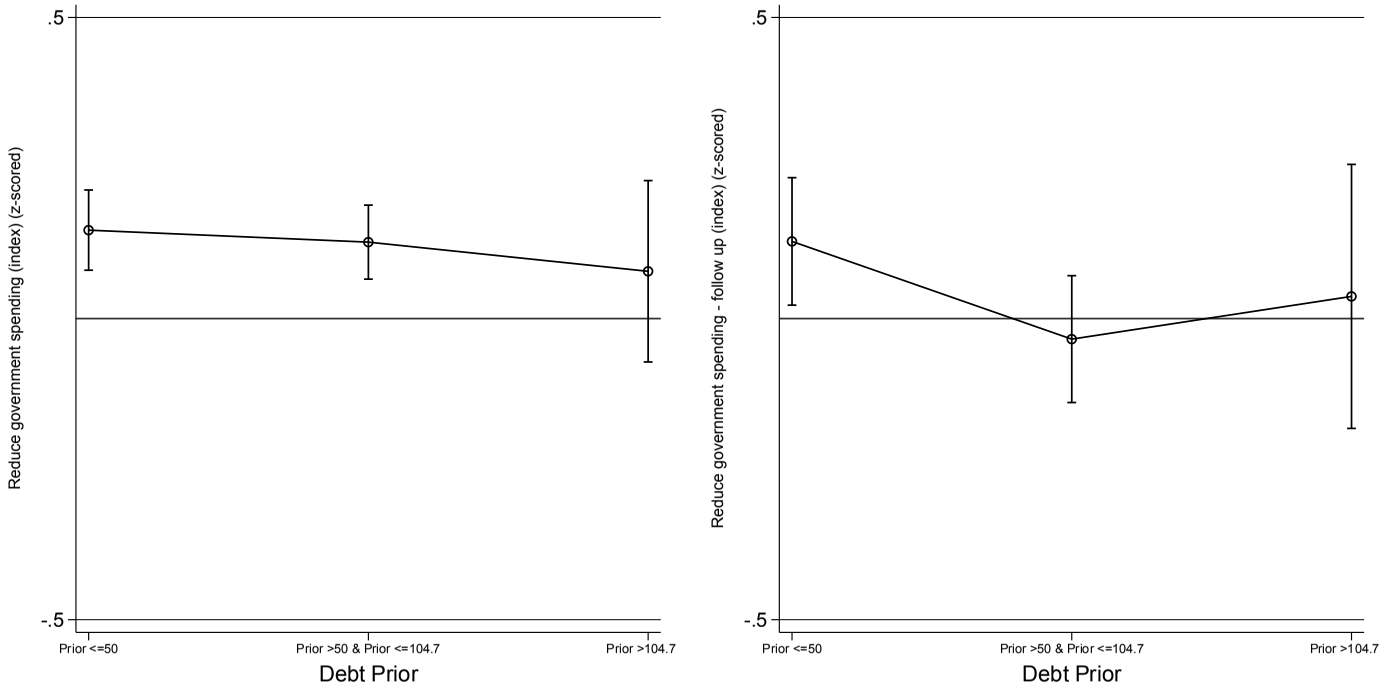
[17] As pre-specified, we winsorize this variable at an estimate of the debt-to-GDP ratio of 200 (i.e. a bias of 95.19) to make sure that our results are not driven by outliers.

Figure 3.4: Heterogeneous effects on views on government debt: by prior beliefs



Notes: This figure describes treatment effects on views on government debt by people's prior beliefs about the debt-to-GDP ratio. The outcome variables are z-scored using the mean and standard deviation in the control group. The figure displays the point estimate of the treatment effects with 90 percent confidence intervals. The figures on the top are based on pooled data from the main experiments on the representative sample and MTurk, while the figures on the bottom are based on the follow-up survey on MTurk. The treatment effect estimates control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample.

Figure 3.5: Heterogeneous effects on attitudes towards government spending (index): by prior beliefs



Notes: This figure describes treatment effects on an index of attitudes towards government spending by people's prior beliefs about the debt-to-GDP ratio. The outcome variable is z-scored using the mean and standard deviation in the control group. The figure displays the point estimate of the treatment effects with 90 percent confidence intervals. The figure on the left is based on pooled data from the main experiments on the representative sample and MTurk, while the figure on the right is based on the follow-up survey on MTurk. The treatment effect estimates control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample.

We also test whether our treatment has heterogeneous effects across demographic groups. Specifically, we estimate specifications including interaction terms of the treatment indicator with a dummy taking value one for Republicans, an indicator variable taking value one if our respondents have at least one child, age, a dummy for high income, and a dummy for respondents with at least a bachelor degree. We find no significant treatment heterogeneity for any of these variables (see Table A.16).

3.5.7 Assessing the external validity of our experimental findings

How externally valid are our experimental estimates? We assess the external validity of the experimental findings by analyzing the correlations of control group respondents' prior beliefs about the debt-to-GDP ratio with their attitudes towards debt reduction, government spending and taxation, their inclination to donate to the Cato Institute and their willingness to sign the petition in favour of a balanced budget rule. In line with our experimental estimates, we show that people who think that the debt-to-GDP ratio is higher are more likely to think that the government should reduce the amount of public debt and that the government should reduce spending (see Table 3.5). Moreover, we find no correlation of this belief with people's attitudes towards taxation, consistent with our experimental findings. However, unlike in the experiment, we find no significant relationship between beliefs about the debt-to-GDP ratio and donations to the Cato institute, even though the coefficient estimate goes into the expected direction. Finally, in contrast to the experimental findings, respondents who think that there is more debt are significantly more likely to sign the petition for the

introduction of a balanced budget-rule.

Taken together, the fact that we find significant correlations between beliefs about the debt-to-GDP ratio and views on debt reduction and government spending, but not views on taxation, and that all correlations have the same sign as our experimental estimates, reassures us of the external validity of our experimental findings. The differences in significance of experimental and correlational estimates for the behavioural outcomes could be due to endogeneity of the OLS results or differential effects of beliefs on policy preferences for the compliant subpopulation of respondents who update their beliefs in response to the information.

3.6 Mechanisms

3.6.1 Why do respondents want to decrease government debt?

In what follows, we examine mechanisms through which our information intervention may increase people's willingness to reduce government debt and to cut government spending.

Intertemporal government budget constraint First, we examine the role of the intertemporal budget constraint of the government. As discussed in Section 3.2, if people form their beliefs in line with the intertemporal government budget constraint, learning that government debt is higher than previously thought should lead them to expect higher taxes or lower government spending in the future. An inclination to smooth the consumption of public goods and taxes over time could then lead them to demand imme-

Table 3.5: Correlates of the demand for government spending and taxation

	Debt Reduction Index	Reduce Total sp.	Increase Total taxes	Petition Index	Donation to Cato Inst.
Est. Debt-to-GDP / 100	0.263**	0.209**	0.097	0.274**	0.093
	(0.107)	(0.091)	(0.101)	(0.127)	(0.134)
Male	-0.118*	-0.008	0.130*	0.065	0.077
	(0.067)	(0.068)	(0.071)	(0.102)	(0.105)
Age	0.005	0.005	-0.004	-0.001	-0.010
	(0.003)	(0.003)	(0.003)	(0.005)	(0.006)
Log(Income)	0.089*	-0.031	0.002	-0.031	0.161**
	(0.048)	(0.047)	(0.050)	(0.072)	(0.078)
Number of children	-0.003	0.064**	-0.032	0.073	0.061
	(0.025)	(0.025)	(0.029)	(0.045)	(0.055)
Employed Full-Time	-0.028	0.112	-0.164	0.058	-0.069
	(0.126)	(0.141)	(0.142)	(0.222)	(0.177)
Employed Part-Time	-0.209	0.135	0.006	-0.076	0.048
	(0.153)	(0.155)	(0.161)	(0.246)	(0.224)
Unemployed	0.053	0.158	-0.336*	0.281	-0.200
	(0.152)	(0.179)	(0.175)	(0.286)	(0.274)
Retired	-0.079	-0.071	0.106	0.466	-0.185
	(0.181)	(0.181)	(0.187)	(0.525)	(0.294)
Student	0.103	0.320	-0.316	0.413	0.385
	(0.208)	(0.215)	(0.224)	(0.417)	(0.468)
High Education	-0.146**	-0.028	0.034	-0.143	-0.056
	(0.070)	(0.070)	(0.071)	(0.104)	(0.095)
Republican	0.408***	0.574***	-0.560***	0.141	0.174
	(0.071)	(0.073)	(0.077)	(0.119)	(0.127)
Observations	812	811	811	384	384

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each column shows one estimation. All estimations are based on the control group in the pooled data from the representative sample and MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

diate cuts in government spending. To shed light on this channel, we ask our respondents whether they think that the current levels of spending and taxation are sustainable, whether they expect changes in spending and taxation for future generations, and how they expect the level of government spending and the tax burden to change between the time of the survey and ten years after the survey.[18]

Table 3.6 provides mixed evidence that our main findings operate through changes in beliefs about future government spending and taxation. We find no significant treatment effects on expectations about spending and taxation in ten years, even though the coefficient estimates go into the expected directions (columns (1) and (2)). Columns (3) and (4) show that people neither significantly update their expectations regarding the levels of spending and taxation that future generations will experience. However, column (5) of Table 3.6 highlights that respondents who receive information about the level of government debt become significantly more likely to think that current public finances are not sustainable. This effect is statistically significant, large in size, robust to adjustment for multiple hypothesis testing, and present for both the MTurk sample and the representative sample. Finally, we find no evidence that treated respondents become more likely to think that it will become more expensive for the government to borrow in the future (column (6)).

Combined, we view these results as suggestive that our findings operate through the perceived intertemporal budget constraint of the government and a consumption-smoothing motive. Treated individuals expect some ad-

[18] We chose the time span of ten years because at this point a new administration will be in office, and we want participants to abstract from specific goals of the current government. Moreover, participants should still be able to form meaningful expectations over this time span.

justment of government spending and taxation to become necessary in the future, while the exact changes they expect as a result of a shift of the perceived level of debt are less clear. One potential explanation for this latter result is that it is very hard for people to predict future government spending or that the time-horizon we picked is too short or too long.

Beliefs about wastage and government efficiency Alternatively, our results could work through reduced trust towards the US government and changes in beliefs about the efficiency of the government. First, after learning that the debt-to-GDP ratio has reached a higher level than they previously thought respondents could become less likely to think that the government can be trusted to do what is right and that the government makes good use of tax money. Second, once people learn about the large amount of government debt, they may update their beliefs about the wastage that occurs in the bureaucratic process. Less trust towards the government and lower perceived efficiency of the public sector could make our respondents more favorable to downsizing the government. Table 3.7 shows that the information did not change people's trust in the government or their beliefs about wastage in the bureaucratic process.

Table 3.6: Expectations about future government spending and taxation and beliefs about fiscal sustainability

	Exp: Increase future taxes	Exp: Decrease future gov. spending	More taxes for future generation	Less gov. spending for future generation	Levels of spend. not sustainable	More expensive to refinance
Panel A: Pooled						
Treatment	0.069	0.048	0.026	0.048	0.148***	0.059
	(0.050)	(0.050)	(0.048)	(0.050)	(0.048)	(0.048)
Adjusted p-value	[0.525]	[0.704]	[0.525]	[0.525]	[0.011]	[0.525]
Observations	1609	1609	1607	1607	1606	1606
Panel B: Rep. Sample						
Treatment	0.115	-0.025	0.044	0.043	0.169**	0.039
	(0.071)	(0.071)	(0.068)	(0.074)	(0.068)	(0.068)
Adjusted p-value	[0.365]	[0.927]	[0.843]	[0.843]	[0.085]	[0.843]
Observations	807	807	805	805	804	804
Panel C: MTurk						
Treatment	0.028	0.108	0.014	0.046	0.120*	0.068
	(0.071)	(0.073)	(0.069)	(0.069)	(0.067)	(0.067)
Adjusted p-value	[1.000]	[0.697]	[1.000]	[1.000]	[0.697]	[0.697]
Observations	802	802	802	802	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table 3.7: Trust in the government and beliefs about government efficiency

	Trust the Gov.	Gov. makes good use of tax money	Gov. is forward-looking	Gov. bureaucracy not efficient	Politicians do not work for public
Panel A: Pooled					
Treatment	0.010	-0.011	-0.050	-0.019	-0.016
	(0.045)	(0.045)	(0.047)	(0.049)	(0.050)
Adjusted p-value	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Observations	1606	1606	1606	1606	1605
Panel B: Rep. Sample					
Treatment	-0.049	-0.069	-0.104	0.063	-0.028
	(0.065)	(0.065)	(0.067)	(0.070)	(0.074)
Adjusted p-value	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Observations	804	804	804	804	803
Panel C: MTurk					
Treatment	0.069	0.039	0.013	-0.089	0.011
	(0.061)	(0.062)	(0.066)	(0.068)	(0.065)
Adjusted p-value	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Observations	802	802	802	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

All in all, the results suggest that our effects operate through the perceived intertemporal government budget constraint and a desire to smooth the consumption of public goods over time rather than through changes in beliefs about wastage and government efficiency.

3.6.2 Robustness

Do the effects operate through genuine changes in beliefs? On the one hand, our treatment could alter people's policy preferences through genuine changes in beliefs as a result of information. On the other hand, the treatment could change people's self-reported policy preferences through channels other than information, such as short-lived emotional responses or priming on the issue that debt is very high. While priming effects should be rather short-lived in nature, effects working through genuine updating of beliefs should persist (Cavallo et al., 2016). In addition, effects working through genuine updating of beliefs should be stronger for individuals with more biased prior beliefs, while this should not be the case for priming effects. Since we find evidence of (i) changes in beliefs about the debt-to-GDP ratio and views on policies that are stronger for individuals with more biased prior beliefs and (ii) persistence of most of our main treatment effects four weeks after the treatment administration it seems likely that our treatment works through information and updating of beliefs rather than priming effects.

Experimenter demand effects We believe that it is unlikely that our results are driven by experimenter demand for at least three reasons. First, we collected data on whether people thought that our survey was politically biased. Overall, 85 percent of respondents felt that the survey was not politically biased. Moreover, our treatment did not shift people's beliefs about whether the survey was politically biased (see Table A.27 in the Appendix). Second, we provide evidence that our effects persist in a four-week follow-up which is much less likely plagued by demand effects (Haaland

and Roth, 2017). Third, de Quidt et al. (2018) find that respondents in online experiments change their behaviour in standard preference measures only very moderately in response to explicit demand manipulations that signal the experimental hypothesis to subjects.

3.7 Conclusion

We provide novel evidence on people's beliefs and preferences regarding government debt. We document several stylized facts using both a representative online panel and an online convenience sample. First, people strongly underestimate the amount of government debt in the US. Second, people's desired amount of government debt is significantly below their estimate of the current debt-to-GDP ratio.

Moreover, we provide new results on the causal effects of beliefs about the debt-to-GDP ratio on people's attitudes towards government spending and taxation. We find strong evidence that respondents who learn that the debt-to-GDP ratio in the US is higher than they thought want the government to reduce the amount of debt and become more supportive of cutting government spending. We even find that people provided with the information donate significantly more money to an NGO lobbying for reducing the size of the government. By contrast, people provided with the information do not alter their views on taxation nor do they become more likely to support a petition in favour of a balanced budget rule. Taken together, our results suggest that learning about the actual amount of government debt lowers people's demand for state-financed public good provision. We show that our treatment effects persist in a four-week follow-up and that respondents in the treatment group have significantly lower biases in beliefs about the

debt-to-GDP ratio four weeks after the treatment was administered.

Our results have several implications. First, our finding that beliefs about the level of government debt causally affect people's policy preferences indicates that voters are forward-looking and take into account economically relevant statistics when forming their views on government policies. Second, our results suggest that support for spending increases could diminish at times when voters update their beliefs about government debt, which could restrict the political feasibility of implementing fiscal stimulus programs during a fiscal crisis. Finally, our finding that voters demand higher levels of spending when they underestimate the level of debt suggests that biased beliefs could contribute to the accumulation of high levels of debt as observed in many industrial countries (Alesina and Passalacqua, 2015).

Appendices of Chapter 3

Summary of the appendices

In Section 3.A we formally show how an increase in people's beliefs about government debt affects their attitudes towards government spending and taxation in a simple two-period model.

In Section 3.B.1 we display summary statistics and provide evidence on the integrity of the randomisation by showing that respondents in the treatment and control groups are balanced in terms of observables. In Section 3.B.2 we provide evidence on correlates of belief updating.

In Section 3.B.3 we replicate our main results from the information experiments without the pre-specified set of control variables. In Section 3.B.4 we present results on the persistence of treatment effects accounting for sample composition by presenting the treatment effects in the main experiment for the subsample of respondents who participated in the follow-up experiment. In Section 3.B.5 we describe heterogeneous treatment effects in response to the information about the debt-to-GDP ratio. We provide additional results on beliefs about political bias, and correlates of beliefs about the debt-to-GDP ratio.

In Section 3.C.1 we illustrate the treatment screen, the actual evolution of the

debt-to-GDP ratio in the US, beliefs about debt-to-GDP, the desired debt-to-GDP ratio, updating in response to the information and heterogeneous response of preferred taxation by prior beliefs. Finally, in Section 3.D.1 we provide additional evidence on the effect of information about the debt-to-GDP ratio on debt-vs. tax-financed spending programs.

3.A Theoretical appendix

In this Section we demonstrate how an increase in the perceived level of government debt affects people's policy preferences in a simple two-period model.

A representative voter who lives for two periods, $t = 1, 2$, has utility over private consumption, c_t, and over consumption of public goods, p_t. We assume that utility is separable between private and public good consumption, and allow for different discount factors for consumption of private and public goods, β and γ:

$$U = u(c_1) + \beta u(c_2) + v(p_1) + \gamma v(p_2)$$

The government can raise revenue by taxing labour income, w_t, in the two periods, which we assume to be exogenous. Given taxes, τ_t, and public good provision by the government, the voter chooses private consumption such as to maximize utility subject to the voter's intertemporal budget constraint. We assume that the voter can borrow and save at the rate $1 + r$:

$$\max_{c_1,c_2} U \qquad \text{s.t.}$$
$$c_1 + \frac{c_2}{1+r} \leq w_1 (1-\tau_1) + \frac{w_2 (1-\tau_2)}{1+r}$$

The voter believes that the government faces the following intertemporal budget constraint:

$$p_1 + \frac{p_2}{1+r} + \frac{\underline{B_3}}{(1+r)^2} \leq \alpha \left[\tau_1 w_1 + \frac{\tau_2 w_2}{1+r} + B_1 \right]$$

where B_1 is the ex-ante net wealth of the government at the beginning of the first period (the negative of government debt), $\underline{B_3}$ is an exogenous lower bound to net wealth of the government at the end of the second period.[19] α lies in the interval $[0, 1]$ and captures the efficiency of the bureaucratic process. We assume that the government can borrow and save at the same rate as the voter, $1 + r$, i.e. that there are no general equilibrium effects on the interest rate.[20]

The voter forms his or her policy preferences by choosing public good provision and taxes in the two periods such as to maximize utility, taking into account the government intertemporal budget constraint and that private consumption will be chosen optimally given taxes and public good provision.

[19] The constraint that debt cannot exceed a certain threshold at the end of the second period captures in a stylized fashion considerations such as constraints to the government's ability to refinance when debt reaches a level that is too high.

[20] One motivation of this is that the government can borrow in international markets.

Assuming log utility for the consumption of private and public goods, $u_t = \log c_t$ and $v_t = \log p_t$, it can be shown that the voter's preferred levels of consumption of private and public goods are given by:

$$
\begin{aligned}
c_1^\star &= \frac{1}{2+\beta+\gamma}\left[w_1 + \frac{w_2}{1+r} + B_1 - \frac{\underline{B_3}}{\alpha\left(1+r\right)^2}\right] \\
c_2^\star &= \frac{\beta\left(1+r\right)}{2+\beta+\gamma}\left[w_1 + \frac{w_2}{1+r} + B_1 - \frac{\underline{B_3}}{\alpha\left(1+r\right)^2}\right] \\
p_1^\star &= \frac{1}{2+\beta+\gamma}\left[\alpha\left(w_1 + \frac{w_2}{1+r} + B_1\right) - \frac{\underline{B_3}}{\left(1+r\right)^2}\right] \\
p_2^\star &= \frac{\gamma\left(1+r\right)}{2+\beta+\gamma}\left[\alpha\left(w_1 + \frac{w_2}{1+r} + B_1\right) - \frac{\underline{B_3}}{\left(1+r\right)^2}\right]
\end{aligned}
$$

Thus, the voter's demand for public spending is increasing in the perceived level of net wealth of the government, B_1, i.e. decreasing in the level of government debt that is inherited in the first period. It is also decreasing in the exogenous lower bound on government net wealth at the end of the second period, B_3, increasing in exogenous labour income in both periods, w_t, and in the efficiency of the government, α.

If the perceived efficiency of the government, α, positively depends on the perceived level of government net wealth, B_1, this will amplify the negative effect of updating beliefs about the level of debt on the voter's demand for public spending:

$$
\frac{\delta \mathbf{p_t^\star}}{\delta \mathbf{B_1}} = \left.\frac{\delta p_t^\star}{\delta B_1}\right|_{\Delta\alpha=0} + \frac{\delta p_t^\star}{\delta \alpha}\frac{\delta \alpha}{\delta B_1} > 0
$$

The net present value of the total tax revenue raised by the government is given by:

$$\tau_1 w_1 + \frac{\tau_2 w_2}{1+r} = \frac{1+\gamma}{2+\beta+\gamma}\left[w_1 + \frac{w_2}{1+r}\right] - \frac{1+\beta}{2+\beta+\gamma}B_1 + \frac{1+\beta}{\alpha\left(2+\beta+\gamma\right)}\frac{\underline{B_3}}{\left(1+r\right)^2}$$

The specific timing of taxes is indeterminate in this model. However, the net present value of taxes increases in the level of government debt at the beginning of the first period. Moreover, in this model a reduced perceived efficiency of the government, $\Delta\alpha < 0$, leads to an increase in total tax revenue collected. Intuitively, if the government works less efficiently, a higher level of taxes will be required for the government to respect the exogenous upper bound on government debt at the beginning of the third period, $-\underline{B}_3$.[21]

Taken together, in a simple two-period model with a representative voter who has log utility over the consumption of public and private goods, an increase in the perceived level of government debt leads to an immediate reduction in the preferred level of government spending. In addition, there is an increase in the net present value of total tax revenue collected. If voters update their beliefs about the efficiency of the government upon learning that government debt is higher than they thought, then this reinforces both the negative effect on the demand for government spending and the positive effect on the net present value of total taxes.

[21] The efficiency of the government affects the relative price of public good consumption, which should lead to both income and substitution effects. Assuming log utility these effects cancel out. The only channel through which the perceived efficiency of the government affects optimal public good provision and taxes is that it makes it more or less difficult to achieve the exogenous lower bound on government net wealth at the end of the second period.

3.B Additional Tables

3.B.1 Summary statistics and balance

Table A.1: Summary statistics: Representative online panel

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.45	0.50	0.00	0.00	1.00	813
Income	62487.70	49004.72	62500.00	0.00	250000.00	813
Age	42.32	15.69	35.00	21.00	70.00	813
Any Children	0.61	0.49	1.00	0.00	1.00	813
Full-time Employed	0.40	0.49	0.00	0.00	1.00	813
Part-time Employed	0.10	0.30	0.00	0.00	1.00	813
Unemployed	0.08	0.27	0.00	0.00	1.00	813
At Least Bachelor	0.40	0.49	0.00	0.00	1.00	813
Republican	0.36	0.48	0.00	0.00	1.00	813
Prior Belief About Debt-to-GDP Ratio	64.78	32.73	60.00	0.00	200.00	813

Table A.2: Summary statistics: MTurk experiment

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.56	0.50	1.00	0.00	1.00	802
Income	58513.09	39100.10	62500.00	0.00	250000.00	802
Age	24.76	11.42	21.00	1.00	57.00	802
Any Children	0.44	0.50	0.00	0.00	1.00	802
Full-time Employed	0.63	0.48	1.00	0.00	1.00	802
Part-time Employed	0.14	0.34	0.00	0.00	1.00	802
Unemployed	0.07	0.26	0.00	0.00	1.00	802
At Least Bachelor	0.51	0.50	1.00	0.00	1.00	802
Republican	0.28	0.45	0.00	0.00	1.00	802
Prior Belief About Debt-to-GDP Ratio	64.28	36.97	55.10	0.00	200.00	802

Table A.3: Summary statistics: MTurk follow-up survey

	Mean	SD	Median	Min.	Max.	Obs.
Male	0.57	0.50	1.00	0.00	1.00	594
Income	58059.76	38875.32	62500.00	0.00	250000.00	594
Age	25.56	11.45	21.00	1.00	57.00	594
Any Children	0.45	0.50	0.00	0.00	1.00	594
Full-time Employed	0.61	0.49	1.00	0.00	1.00	594
Part-time Employed	0.14	0.35	0.00	0.00	1.00	594
Unemployed	0.08	0.27	0.00	0.00	1.00	594
At Least Bachelor	0.52	0.50	1.00	0.00	1.00	594
Republican	0.27	0.45	0.00	0.00	1.00	594
Prior Belief About Debt-to-GDP Ratio	65.15	37.81	56.50	0.00	200.00	594

Table A.4: Characteristics of the representative sample compared to the American Community Survey (ACS)

	Mean: Rep. Online sample	Mean: ACS
Female	0.55	0.51
Age 18-24	0.113	0.126
Age 25-34	0.211	0.177
Age 35-44	0.187	0.165
Age 45-54	0.169	0.174
Age 55-64	0.189	0.165
Age 65 and older	0.131	0.193
Northeast	0.187	0.179
Midwest	0.225	0.211
South	0.370	0.375
West	0.218	0.24
Total household income	62,487	84,568

Notes: This table summarizes the characteristics of our sample from the representative online panel as well as the characteristics of the 2015 American Community Survey.

Table A.5: Balance: Representative online panel

	Treatment	Control	P-value(Treatment - Control)	Observations
Prior Belief About Debt-to-GDP Ratio	63.98	65.50	0.507	813
Male	0.44	0.45	0.780	813
Age	42.10	42.51	0.711	813
Log(Income)	10.54	10.63	0.414	813
Number of Children	1.26	1.29	0.734	803
Unemployed	0.07	0.09	0.398	813
Part-time Employed	0.10	0.10	0.853	813
Full-time Employed	0.41	0.40	0.660	813
Retired	0.18	0.20	0.395	813
Student	0.05	0.05	0.713	813
Other Employment Status	0.09	0.09	0.898	813
At Least Bachelor	0.41	0.40	0.819	813
Republican	0.30	0.41	0.002	813

The p-value of a joint F-test when regressing the treatment dummy on all covariates is 0.2431.

Table A.6: Balance: MTurk experiment

	Treatment	Control	P-value(Treatment - Control)	Observations
Prior Belief About Debt-to-GDP Ratio	62.71	65.99	0.211	802
Male	0.59	0.53	0.120	802
Age	25.37	24.08	0.108	802
Log(Income)	10.66	10.52	0.199	802
Number of Children	0.90	0.89	0.892	802
Unemployed	0.08	0.06	0.191	802
Part-time Employed	0.14	0.13	0.583	802
Full-time Employed	0.61	0.66	0.135	802
Retired	0.03	0.02	0.100	802
Student	0.03	0.03	0.858	802
Other Employment Status	0.06	0.06	0.768	802
At Least Bachelor	0.50	0.52	0.552	802
Republican	0.28	0.28	0.974	802

The p-value of a joint F-test when regressing the treatment dummy on all covariates is 0.1311.

Table A.7: Balance: MTurk follow-up

	Treatment	Control	P-value(Treatment - Control)	Observations
Prior Belief About Debt-to-GDP Ratio	63.99	66.49	0.424	594
Male	0.59	0.54	0.278	594
Age	25.85	25.23	0.509	594
Log(Income)	10.65	10.55	0.413	594
Number of Children	0.88	0.91	0.728	594
Unemployed	0.09	0.07	0.500	594
Part-time Employed	0.16	0.12	0.206	594
Full-time Employed	0.57	0.66	0.018	594
Retired	0.04	0.02	0.128	594
Student	0.03	0.02	0.212	594
Other Employment Status	0.06	0.07	0.638	594
At Least Bachelor	0.50	0.53	0.476	594
Republican	0.26	0.28	0.581	594

The p-value of a joint F-test when regressing the treatment dummy on all covariates is 0.3326.

3.B.2 Correlates of belief updating

Table A.8: Correlates of belief updating

	Change in est. debt-to-GDP ratio
Est. Debt-to-GDP / 100	-73.149***
	(7.084)
Male	8.192**
	(3.941)
Age	0.182
	(0.173)
Log(Income)	1.423
	(2.839)
Number of children	-1.332
	(1.847)
Employed Full-Time	-12.491
	(9.969)
Employed Part-Time	-8.534
	(11.052)
Unemployed	-7.053
	(11.579)
Retired	-10.471
	(13.000)
Student	-22.512*
	(11.999)
High Education	2.208
	(4.195)
Republican	-0.608
	(4.805)
Observations	316

Notes: The estimation is based on the treatment group in the MTurk subsample that completed the follow-up. The specification controls for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). Robust standard errors are in parentheses.

3.B.3 Main results without controls

Table A.9: Views on government debt: Without controls

	There is too much debt	Gov. should reduce debt	Index
Panel A: Pooled			
Treatment	0.284***	0.221***	0.241***
	(0.045)	(0.046)	(0.044)
Adjusted p-value	[0.001]	[0.001]	
Observations	1614	1614	1614
Panel B: Rep. Sample			
Treatment	0.295***	0.245***	0.260***
	(0.063)	(0.063)	(0.061)
Adjusted p-value	[0.001]	[0.001]	
Observations	812	812	812
Panel C: MTurk			
Treatment	0.272***	0.197***	0.221***
	(0.066)	(0.068)	(0.065)
Adjusted p-value	[0.001]	[0.001]	
Observations	802	802	802
Panel D: Follow-up Sample			
Treatment	0.117	0.138*	0.128*
	(0.080)	(0.076)	(0.074)
Adjusted p-value	[0.158]	[0.158]	
Observations	599	599	599

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. Panel D present results from the follow-up experiment on MTurk. The estimations on the pooled sample control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table A.10: Attitudes towards government spending: Without controls

	Reduce Overall Sp.	Reduce Defense Sp.	Reduce Infrastr. Sp.	Reduce Schooling Sp.	Reduce Social Sec. Sp.	Reduce Social Ins. Sp.	Reduce Health Sp.	Reduce Environm. Sp.	Index
Panel A: Pooled									
Treatment	0.127**	0.052	0.109**	0.141***	0.132***	0.087*	0.103**	0.082*	0.104***
	(0.050)	(0.050)	(0.050)	(0.050)	(0.049)	(0.049)	(0.050)	(0.049)	(0.029)
Adjusted p-value	[0.027]	[0.115]	[0.040]	[0.027]	[0.027]	[0.065]	[0.041]	[0.065]	
Observations	1613	1613	1613	1613	1613	1613	1613	1613	1613
Panel B: Rep. Sample									
Treatment	0.114	0.103	0.187***	0.131*	0.161**	0.026	0.105	0.067	0.112***
	(0.072)	(0.070)	(0.072)	(0.072)	(0.070)	(0.071)	(0.072)	(0.070)	(0.042)
Adjusted p-value	[0.168]	[0.168]	[0.079]	[0.166]	[0.084]	[0.238]	[0.168]	[0.238]	
Observations	811	811	811	811	811	811	811	811	811
Panel C: MTurk									
Treatment	0.141**	0.001	0.030	0.152**	0.102	0.149**	0.101	0.096	0.096**
	(0.068)	(0.070)	(0.071)	(0.069)	(0.067)	(0.068)	(0.070)	(0.069)	(0.040)
Adjusted p-value	[0.114]	[0.327]	[0.289]	[0.114]	[0.163]	[0.114]	[0.163]	[0.163]	
Observations	802	802	802	802	802	802	802	802	802
Panel D: Follow-up Sample									
Treatment	0.105	-0.012	0.014	0.007	0.054	0.040	0.047	-0.073	0.023
	(0.077)	(0.083)	(0.083)	(0.079)	(0.083)	(0.082)	(0.081)	(0.081)	(0.048)
Adjusted p-value	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	
Observations	599	597	597	597	597	597	597	597	597

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. Panel D present results from the follow-up experiment on MTurk. The estimations on the pooled sample control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table A.11: Attitudes towards taxation: Without controls

	Increase overall amount of taxes	Increase income tax: top 10	Increase income tax: next 40	Increase income tax: bottom 50	Introduce wealth tax	Increase estate tax	Index
Panel A: Pooled							
Treatment	0.128***	0.060	0.037	0.030	-0.007	0.102**	0.043
	(0.049)	(0.050)	(0.049)	(0.051)	(0.050)	(0.050)	(0.030)
Adjusted p-value	[0.059]	[0.437]	[0.787]	[0.787]	[0.841]	[0.113]	
Observations	1613	1612	1612	1612	1611	1610	1610
Panel B: Rep. Sample							
Treatment	0.102	0.044	0.011	-0.059	0.092	0.193***	0.040
	(0.071)	(0.073)	(0.070)	(0.073)	(0.069)	(0.069)	(0.042)
Adjusted p-value	[0.437]	[0.574]	[0.771]	[0.574]	[0.437]	[0.032]	
Observations	811	810	810	810	809	809	809
Panel C: MTurk							
Treatment	0.154**	0.078	0.064	0.120*	-0.106	0.010	0.046
	(0.068)	(0.069)	(0.069)	(0.070)	(0.072)	(0.071)	(0.043)
Adjusted p-value	[0.171]	[0.358]	[0.400]	[0.282]	[0.314]	[0.634]	
Observations	802	802	802	802	802	801	801
Panel D: Follow-up Sample							
Treatment	0.089	-0.102	0.045	0.155*	0.122	-0.057	0.041
	(0.080)	(0.086)	(0.079)	(0.079)	(0.084)	(0.083)	(0.041)
Adjusted p-value	[0.499]	[0.499]	[0.665]	[0.439]	[0.499]	[0.665]	
Observations	597	597	597	597	596	593	593

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. Panel D present results from the follow-up experiment on MTurk. The estimations on the pooled sample control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table A.12: behavioural measures: Without controls

	Petition in favor of a balanced budget rule				Donation to
	Want to sign	Report: Signed	Index: Self-reports	Actual signatures	Cato Institute
	(1)	(2)	(3)	(4)	(5)
Treatment	0.0386	-0.0105	0.00554	0.031	0.141*
	(0.032)	(0.027)	(0.027)	(0.019)	(0.077)
Control group mean	0	0	0	0.06	0
Observations	802	802	802	802	802

Notes: The outcome variables in columns 1, 2, 3 and 5 are z-scored using the mean and standard deviation in the control group. All estimations are based on the MTurk sample. Robust standard errors are in parentheses.

3.B.4 Persistence of the effect accounting for sample composition

Table A.13: Views on government debt: Sample composition effects

	There is too much debt	Gov. should reduce debt	Index
Panel A: MTurk Main			
Treatment	0.301***	0.218***	0.245***
	(0.063)	(0.065)	(0.062)
Adjusted p-value	[0.001]	[0.001]	
Observations	802	802	802
Panel B: MTurk Main (follow-up sample)			
Treatment	0.378***	0.287***	0.317***
	(0.074)	(0.075)	(0.071)
Adjusted p-value	[0.001]	[0.001]	
Observations	599	599	599
Panel C: MTurk Follow-up			
Treatment	0.159**	0.177**	0.168**
	(0.075)	(0.073)	(0.070)
Adjusted p-value	[0.033]	[0.033]	
Observations	599	599	599

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the MTurk sample from the main Experiment, Panel B shows estimations on the results from the main experiment from the MTurk sample that completed the follow-up and Panel C shows results from the follow-up experiment. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table A.14: Attitudes towards government spending: Sample composition effects

	Reduce Overall Sp.	Reduce Defense Sp.	Reduce Infrastr. Sp.	Reduce Schooling Sp.	Reduce Social Sec. Sp.	Reduce Social Ins. Sp.	Reduce Health Sp.	Reduce Environm. Sp.	Index
Panel A: MTurk Main									
Treatment	0.163**	-0.004	0.052	0.131**	0.100	0.153**	0.094	0.093	0.098***
	(0.063)	(0.063)	(0.068)	(0.064)	(0.063)	(0.063)	(0.063)	(0.060)	(0.036)
Adjusted p-value	[0.067]	[0.313]	[0.222]	[0.086]	[0.128]	[0.067]	[0.128]	[0.128]	
Observations	802	802	802	802	802	802	802	802	802
Panel B: MTurk Main (follow-up sample)									
Treatment	0.152**	-0.043	0.091	0.168**	0.102	0.142*	0.137*	0.101	0.107***
	(0.070)	(0.071)	(0.078)	(0.075)	(0.072)	(0.074)	(0.072)	(0.069)	(0.041)
Adjusted p-value	[0.135]	[0.262]	[0.163]	[0.135]	[0.135]	[0.135]	[0.135]	[0.135]	
Observations	599	597	597	597	597	597	597	597	597
Panel C: MTurk Follow-up									
Treatment	0.156**	-0.025	0.047	0.025	0.077	0.085	0.073	-0.042	0.050
	(0.073)	(0.074)	(0.082)	(0.074)	(0.079)	(0.074)	(0.071)	(0.070)	(0.043)
Adjusted p-value	[0.339]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	
Observations	599	597	597	597	597	597	597	597	597

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the MTurk sample from the main Experiment, Panel B shows estimations on the results from the main experiment from the MTurk sample that completed the follow-up and Panel C shows results from the follow-up experiment. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

Table A.15: Attitudes towards taxation: Sample composition effects

	Increase overall amount of taxes	Increase income tax: top 10	Increase income tax: next 40	Increase income tax: bottom 50	Introduce wealth tax	Increase estate tax	Index
Panel A: MTurk Main							
Treatment	0.131**	0.067	0.050	0.135**	-0.096	-0.003	0.047
	(0.062)	(0.064)	(0.066)	(0.068)	(0.069)	(0.067)	(0.039)
Adjusted p-value	[0.171]	[0.422]	[0.489]	[0.171]	[0.280]	[0.803]	
Observations	802	802	802	802	802	801	801
Panel B: MTurk Main (follow-up sample)							
Treatment	0.155**	0.000	0.101	0.160**	-0.105	0.010	0.054
	(0.071)	(0.073)	(0.077)	(0.080)	(0.081)	(0.079)	(0.046)
Adjusted p-value	[0.154]	[0.499]	[0.246]	[0.154]	[0.246]	[0.499]	
Observations	597	597	597	597	596	593	593
Panel C: MTurk Follow-up							
Treatment	0.030	-0.138*	0.012	0.151*	0.144*	-0.085	0.017
	(0.074)	(0.082)	(0.075)	(0.079)	(0.080)	(0.079)	(0.039)
Adjusted p-value	[0.691]	[0.222]	[0.733]	[0.222]	[0.222]	[0.268]	
Observations	597	597	597	597	596	593	593

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

3.B.5 Heterogeneous effects

One could imagine that people's political affiliation plays an important role in shaping people's response to our information treatment. In particular, Republicans already have a very strong preference for downsizing the government, which reduces the available variation to change their preferences. However, it is also possible that Republicans could use the high levels of debt as an excuse to demand further decreases in government spending. We find little evidence that the responses to our treatment depend on people's party affiliation (Table A.16). It is possible that people with different levels of educational attainment respond differently to information about the debt-to-GDP ratio. People with more education might respond less to the information treatment as they are less biased about the true statistic than are people with low levels of education. But, they might respond more strongly to the treatment as they are more numerate. There is little evidence that the strength of the treatment effect depends on education. We do not find strong evidence for heterogeneity according to income (Table A.16). We also look at heterogeneous treatment effects by age. We expect young individuals to respond more strongly to the treatment as they are more likely to see higher taxes and lower government spending in the future which might become necessary in order to reduce government debt. We find no significant heterogeneity by age (Table A.16). To test whether concerns for future generations moderate the response to our treatment, we also examine heterogeneity by a dummy variable taking value one if the respondent reports having at least one child. There is no significant heterogeneous response along this dimension except for people's perception that debt is too high, which is stronger for respondents with children (Table A.16).

Table A.16: Heterogeneous treatment effects

	Debt Reduction Index	Reduce Overall Sp.	Increase overall Amount of taxes	Petition Index	Donation to Cato Inst.
Treatment × Est. Debt-to-GDP / 100	-0.102 (0.138)	-0.094 (0.140)	-0.078 (0.144)	-0.252 (0.181)	-0.076 (0.192)
Treatment × Republican	-0.093 (0.087)	-0.145 (0.094)	-0.022 (0.102)	0.127 (0.163)	0.109 (0.183)
Treatment × High Education	0.009 (0.089)	-0.072 (0.099)	0.091 (0.098)	0.214 (0.141)	-0.093 (0.155)
Treatment × High Income	0.079 (0.089)	0.175* (0.099)	0.025 (0.100)	0.169 (0.149)	-0.037 (0.161)
Treatment × Age	-0.003 (0.003)	-0.003 (0.003)	-0.000 (0.003)	0.004 (0.007)	0.017** (0.008)
Treatment × Children	0.239** (0.093)	0.115 (0.100)	-0.025 (0.103)	-0.069 (0.154)	-0.121 (0.164)
Treatment	0.198 (0.126)	0.159 (0.136)	0.037 (0.142)	-0.397* (0.216)	-0.220 (0.251)
Observations	1614	1613	1613	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each column shows one estimation and every estimation is done on the pooled sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.17: Heterogeneous treatment effects: Low debt-to-GDP ratio

	Debt Reduction Index	Reduce Overall Sp.	Increase overall Amount of taxes	Petition Index	Donation to Cato Inst.
Treatment ×	-0.015	-0.043	0.281**	0.212	0.334
Low Est. Debt-to-GDP	(0.128)	(0.131)	(0.136)	(0.185)	(0.208)
Treatment ×	-0.090	-0.142	-0.033	0.117	0.093
Republican	(0.087)	(0.094)	(0.102)	(0.163)	(0.183)
Treatment ×	0.009	-0.072	0.092	0.214	-0.091
High Education	(0.089)	(0.099)	(0.098)	(0.141)	(0.155)
Treatment ×	0.082	0.177*	0.025	0.170	-0.021
High Income	(0.089)	(0.098)	(0.099)	(0.148)	(0.161)
Treatment ×	-0.003	-0.003	-0.000	0.003	0.018**
Age	(0.003)	(0.003)	(0.003)	(0.006)	(0.008)
Treatment ×	0.242***	0.119	-0.034	-0.069	-0.137
Children	(0.093)	(0.100)	(0.103)	(0.154)	(0.164)
Treatment	0.260*	0.241	-0.170	-0.456*	-0.474
	(0.152)	(0.164)	(0.173)	(0.251)	(0.293)
Observations	1614	1613	1613	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each column shows one estimation and every estimation is done on the pooled sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.18: Heterogeneous treatment effects: Below median debt-to-GDP ratio

	Debt Reduction Index	Reduce Overall Sp.	Increase overall Amount of taxes	Petition Index	Donation to Cato Inst.
Treatment ×	0.159*	0.089	-0.025	0.216	-0.025
Below Median Est. Debt-to-GDP	(0.085)	(0.093)	(0.093)	(0.140)	(0.149)
Treatment ×	-0.083	-0.139	-0.024	0.138	0.104
Republican	(0.087)	(0.094)	(0.103)	(0.163)	(0.182)
Treatment ×	0.011	-0.071	0.089	0.209	-0.094
High Education	(0.089)	(0.099)	(0.098)	(0.141)	(0.155)
Treatment ×	0.080	0.175*	0.024	0.165	-0.042
High Income	(0.089)	(0.099)	(0.100)	(0.149)	(0.161)
Treatment ×	-0.003	-0.003	-0.000	0.004	0.017**
Age	(0.003)	(0.003)	(0.003)	(0.007)	(0.008)
Treatment ×	0.238**	0.114	-0.022	-0.060	-0.115
Children	(0.092)	(0.100)	(0.103)	(0.154)	(0.164)
Treatment	0.146	0.148	0.088	-0.413*	-0.168
	(0.122)	(0.135)	(0.138)	(0.217)	(0.248)
Observations	1614	1613	1613	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each column shows one estimation and every estimation is done on the pooled sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.19: Heterogeneous treatment effects: Spending categories

	Reduce Defense Sp.	Reduce Infrastr. Sp.	Reduce Schooling Sp.	Reduce Social Sec. Sp.	Reduce Social Ins. Sp.	Reduce Health Sp.	Reduce Environm. Sp.
Treatment ×	-0.233*	-0.142	-0.039	-0.076	0.008	0.228	0.034
Est. Debt-to-GDP / 100	(0.138)	(0.152)	(0.140)	(0.125)	(0.124)	(0.139)	(0.133)
Treatment ×	-0.153	-0.122	0.014	-0.174*	-0.140	0.074	0.052
Republican	(0.099)	(0.107)	(0.104)	(0.102)	(0.099)	(0.104)	(0.096)
Treatment ×	-0.165*	-0.046	-0.123	-0.062	-0.070	0.059	0.002
High Education	(0.094)	(0.101)	(0.097)	(0.094)	(0.093)	(0.095)	(0.092)
Treatment ×	0.280***	-0.070	-0.152	-0.100	-0.052	-0.101	-0.083
High Income	(0.096)	(0.102)	(0.099)	(0.095)	(0.095)	(0.097)	(0.094)
Treatment ×	-0.002	-0.002	0.002	0.001	-0.001	-0.001	-0.003
Age	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Treatment ×	-0.078	0.062	0.113	-0.009	0.176*	0.146	0.105
Children	(0.099)	(0.104)	(0.099)	(0.099)	(0.097)	(0.100)	(0.096)
Treatment	0.006	0.200	0.165	0.236*	0.178	0.214	0.204
	(0.139)	(0.150)	(0.146)	(0.142)	(0.133)	(0.142)	(0.138)
Observations	1613	1613	1613	1613	1613	1613	1613

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each column shows one estimation and every estimation is done on the pooled sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.20: Heterogeneous treatment effects: Tax categories

	Increase income tax: top 10	Increase income tax: next 40	Increase income tax: bottom 50	Introduce wealth tax	Increase estate tax
Treatment × Est. Debt-to-GDP / 100	0.106 (0.145)	-0.076 (0.141)	0.052 (0.150)	0.066 (0.140)	-0.146 (0.146)
Treatment × Republican	-0.107 (0.113)	0.046 (0.111)	-0.094 (0.108)	-0.024 (0.111)	-0.047 (0.105)
Treatment × High Education	0.031 (0.097)	0.108 (0.099)	0.162 (0.103)	0.027 (0.100)	-0.091 (0.098)
Treatment × High Income	0.043 (0.099)	-0.062 (0.100)	-0.017 (0.104)	0.072 (0.101)	0.043 (0.100)
Treatment × Age	-0.006* (0.003)	-0.000 (0.003)	-0.002 (0.003)	0.002 (0.003)	0.004 (0.003)
Treatment × Children	0.042 (0.103)	-0.036 (0.103)	-0.023 (0.108)	-0.009 (0.103)	-0.089 (0.103)
Treatment	0.226 (0.144)	-0.028 (0.139)	0.099 (0.157)	-0.114 (0.143)	-0.044 (0.145)
Observations	1612	1612	1612	1611	1610

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each column shows one estimation and every estimation is done on the pooled sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.21: Heterogeneous treatment effects: Separate regressions for demographics

	Debt Reduction Index	Reduce Overall Sp.	Increase overall Amount of taxes	Petition Index	Donation to Cato Inst.
Panel A: Republican					
Treatment × Republican	-0.050	-0.119	-0.027	0.127	0.106
	(0.085)	(0.093)	(0.099)	(0.158)	(0.178)
Treatment	0.296***	0.215***	0.100*	-0.023	0.122
	(0.056)	(0.060)	(0.057)	(0.077)	(0.088)
Observations	1614	1613	1613	802	802
Panel B: Education					
Treatment × High Education	0.026	-0.026	0.101	0.238*	-0.065
	(0.087)	(0.093)	(0.093)	(0.135)	(0.151)
Treatment	0.268***	0.189***	0.045	-0.110	0.186
	(0.056)	(0.061)	(0.064)	(0.101)	(0.116)
Observations	1614	1613	1613	802	802
Panel C: Income					
Treatment × High Income	0.113	0.168*	0.045	0.206	-0.065
	(0.087)	(0.093)	(0.094)	(0.139)	(0.154)
Treatment	0.220***	0.087	0.068	-0.108	0.180
	(0.065)	(0.069)	(0.070)	(0.106)	(0.114)
Observations	1614	1613	1613	802	802
Panel D: Age					
Treatment × Age	-0.001	-0.003	-0.001	0.003	0.015**
	(0.003)	(0.003)	(0.003)	(0.006)	(0.008)
Treatment	0.326***	0.261**	0.120	-0.061	-0.222
	(0.098)	(0.102)	(0.102)	(0.162)	(0.213)
Observations	1614	1613	1613	802	802
Panel E: Children					
Treatment × Children	0.203**	0.087	-0.031	0.019	0.021
	(0.086)	(0.093)	(0.093)	(0.141)	(0.153)
Treatment	0.173***	0.128*	0.110	0.001	0.146
	(0.063)	(0.068)	(0.067)	(0.086)	(0.104)
Observations	1614	1613	1613	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each Panel shows results for a different dimension of heterogeneity. All estimations are based on the pooled data from the representative sample and from MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.22: Heterogeneous treatment effects: Separate regressions for measures of the bias

	Debt Reduction Index	Reduce Overall Sp.	Increase overall Amount of taxes	Petition Index	Donation to Cato Inst.
Panel A: Est. Debt-to-GDP (continuous)					
Treatment ×	-0.001	-0.001	-0.000	-0.002	-0.000
Est. Debt-to-GDP	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Treatment	0.209***	0.100	0.101	-0.079	0.135
	(0.072)	(0.078)	(0.079)	(0.102)	(0.113)
Observations	1614	1613	1613	802	802
Panel B: Belief (low debt-to-GDP)					
Treatment ×	0.001	-0.039	0.262*	0.207	0.330
Low Debt-to-GDP	(0.130)	(0.138)	(0.142)	(0.186)	(0.206)
Treatment	0.260**	0.184	-0.102	-0.160	-0.122
	(0.121)	(0.127)	(0.132)	(0.170)	(0.190)
Observations	1614	1613	1613	802	802
Panel C: Below Median Est. Debt-to-GDP					
Treatment ×	0.178**	0.120	-0.051	0.183	-0.061
Below Median Est. Debt-to-GDP	(0.086)	(0.095)	(0.096)	(0.138)	(0.153)
Treatment	0.168***	0.088	0.145**	-0.084	0.183
	(0.059)	(0.067)	(0.070)	(0.104)	(0.118)
Observations	1614	1613	1613	802	802
Panel D: Underestimators					
Treatment ×	0.044	0.057	0.097	-0.100	0.223
Underestimate	(0.163)	(0.182)	(0.192)	(0.227)	(0.234)
Treatment	0.217	0.097	0.032	0.097	-0.057
	(0.157)	(0.175)	(0.185)	(0.215)	(0.221)
Observations	1614	1613	1613	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each Panel shows results for a different dimension of heterogeneity. All estimations are based on the pooled data from the representative sample and from MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.23: Heterogeneous treatment effects: Separate regressions for demographics - spending categories

	Reduce Defense Sp.	Reduce Infrastr. Sp.	Reduce Schooling Sp.	Reduce Social Sec. Sp.	Reduce Social Ins. Sp.	Reduce Health Sp.	Reduce Environm. Sp.
Panel A: Republican							
Treatment × Republican	-0.156	-0.117	0.041	-0.174*	-0.106	0.093	0.060
	(0.097)	(0.104)	(0.101)	(0.098)	(0.096)	(0.101)	(0.093)
Treatment	0.061	0.180***	0.173***	0.230***	0.185***	0.126**	0.118**
	(0.055)	(0.060)	(0.055)	(0.055)	(0.055)	(0.055)	(0.054)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel B: Education							
Treatment × High Education	-0.085	-0.066	-0.170*	-0.091	-0.093	0.026	-0.022
	(0.091)	(0.098)	(0.092)	(0.091)	(0.090)	(0.092)	(0.088)
Treatment	0.050	0.172**	0.264***	0.216***	0.195***	0.144**	0.147**
	(0.062)	(0.068)	(0.064)	(0.063)	(0.063)	(0.064)	(0.062)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel C: Income							
Treatment × High Income	0.220**	-0.079	-0.173*	-0.130	-0.055	-0.059	-0.061
	(0.091)	(0.098)	(0.093)	(0.092)	(0.090)	(0.092)	(0.089)
Treatment	-0.106	0.184**	0.279***	0.245***	0.182***	0.188***	0.169**
	(0.066)	(0.072)	(0.069)	(0.066)	(0.067)	(0.068)	(0.067)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel D: Age							
Treatment × Age	-0.004	-0.001	0.004	0.001	0.001	0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Treatment	0.137	0.185*	0.048	0.135	0.114	0.135	0.179*
	(0.103)	(0.108)	(0.104)	(0.108)	(0.102)	(0.102)	(0.099)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel E: Children							
Treatment × Children	-0.082	0.018	0.124	-0.037	0.141	0.128	0.078
	(0.091)	(0.098)	(0.093)	(0.092)	(0.089)	(0.091)	(0.088)
Treatment	0.054	0.130*	0.123*	0.190***	0.070	0.082	0.093
	(0.066)	(0.068)	(0.067)	(0.065)	(0.064)	(0.063)	(0.063)
Observations	1613	1613	1613	1613	1613	1613	1613

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each Panel shows results for a different dimension of heterogeneity. All estimations are based on the pooled data from the representative sample and from MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.24: Heterogeneous treatment effects: Separate regressions for measures of the bias - spending categories

	Reduce Defense Sp.	Reduce Infrastr. Sp.	Reduce Schooling Sp.	Reduce Social Sec. Sp.	Reduce Social Ins. Sp.	Reduce Health Sp.	Reduce Environm. Sp.
Panel A: Est. Debt-to-GDP (continuous)							
Treatment ×	-0.002	-0.002	-0.001	-0.001	-0.000	0.002	-0.000
Est. Debt-to-GDP	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Treatment	-0.030	0.056	0.115	0.109	0.101	0.192**	0.079
	(0.076)	(0.080)	(0.084)	(0.072)	(0.076)	(0.080)	(0.078)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel B: Belief (low debt-to-GDP)							
Treatment ×	0.116	0.129	0.169	0.188	-0.027	-0.164	0.068
Low Debt-to-GDP	(0.139)	(0.146)	(0.153)	(0.137)	(0.144)	(0.143)	(0.141)
Treatment	-0.051	0.014	0.007	-0.010	0.136	0.256*	0.034
	(0.129)	(0.136)	(0.144)	(0.127)	(0.135)	(0.133)	(0.131)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel C: Below Median Est. Debt-to-GDP							
Treatment ×	0.040	-0.053	0.039	0.005	0.073	-0.043	-0.013
Below Median Est. Debt-to-GDP	(0.097)	(0.099)	(0.098)	(0.094)	(0.096)	(0.098)	(0.097)
Treatment	0.027	0.153**	0.127*	0.146**	0.076	0.133*	0.097
	(0.070)	(0.070)	(0.074)	(0.068)	(0.071)	(0.071)	(0.070)
Observations	1613	1613	1613	1613	1613	1613	1613
Panel D: Underestimators							
Treatment ×	0.220	0.067	0.055	0.142	0.003	-0.171	0.036
Underestimate	(0.178)	(0.185)	(0.204)	(0.166)	(0.171)	(0.183)	(0.186)
Treatment	-0.151	0.060	0.097	0.017	0.108	0.270	0.055
	(0.171)	(0.177)	(0.197)	(0.159)	(0.163)	(0.176)	(0.179)
Observations	1613	1613	1613	1613	1613	1613	1613

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each Panel shows results for a different dimension of heterogeneity. All estimations are based on the pooled data from the representative sample and from MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.25: Heterogeneous treatment effects: Separate regressions for demographics - tax categories

	Increase income tax: top 10	Increase income tax: next 40	Increase income tax: bottom 50	Introduce wealth tax	Increase estate tax
Panel A: Republican					
Treatment ×	-0.115	0.031	-0.104	-0.015	-0.046
Republican	(0.111)	(0.108)	(0.107)	(0.109)	(0.103)
Treatment	0.052	-0.008	0.082	-0.035	0.078
	(0.052)	(0.055)	(0.060)	(0.055)	(0.057)
Observations	1612	1612	1612	1611	1610
Panel B: Education					
Treatment ×	0.055	0.089	0.163*	0.043	-0.087
High Education	(0.093)	(0.096)	(0.099)	(0.096)	(0.094)
Treatment	-0.010	-0.039	-0.026	-0.060	0.103
	(0.065)	(0.066)	(0.068)	(0.062)	(0.063)
Observations	1612	1612	1612	1611	1610
Panel C: Income					
Treatment ×	0.063	-0.037	0.020	0.074	-0.004
High Income	(0.094)	(0.097)	(0.099)	(0.096)	(0.095)
Treatment	-0.019	0.021	0.040	-0.079	0.067
	(0.072)	(0.072)	(0.074)	(0.069)	(0.071)
Observations	1612	1612	1612	1611	1610
Panel D: Age					
Treatment ×	-0.006*	-0.001	-0.002	0.002	0.003
Age	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Treatment	0.206*	0.032	0.118	-0.095	-0.033
	(0.107)	(0.105)	(0.113)	(0.107)	(0.107)
Observations	1612	1612	1612	1611	1610
Panel E: Children					
Treatment ×	-0.031	-0.045	-0.065	0.012	-0.050
Children	(0.093)	(0.095)	(0.099)	(0.096)	(0.094)
Treatment	0.029	0.022	0.080	-0.049	0.087
	(0.067)	(0.067)	(0.073)	(0.069)	(0.067)
Observations	1612	1612	1612	1611	1610

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each Panel shows results for a different dimension of heterogeneity. All estimations are based on the pooled data from the representative sample and from MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.26: Heterogeneous treatment effects: Separate regressions for measures of the bias - tax categories

	Increase income tax: top 10	Increase income tax: next 40	Increase income tax: bottom 50	Introduce wealth tax	Increase estate tax
Panel A: Est. Debt-to-GDP (continuous)					
Treatment ×	0.001	-0.000	0.001	0.001	-0.001
Est. Debt-to-GDP	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Treatment	0.108	0.014	0.065	0.032	0.056
	(0.083)	(0.078)	(0.081)	(0.077)	(0.079)
Observations	1612	1612	1612	1611	1610
Panel B: Belief (low debt-to-GDP)					
Treatment ×	-0.186	0.027	0.144	-0.027	0.212
Low Debt-to-GDP	(0.154)	(0.143)	(0.151)	(0.142)	(0.144)
Treatment	0.206	0.005	-0.076	0.008	-0.088
	(0.145)	(0.133)	(0.142)	(0.131)	(0.134)
Observations	1612	1612	1612	1611	1610
Panel C: Below Median Est. Debt-to-GDP					
Treatment ×	-0.125	0.023	-0.152	-0.173*	-0.039
Below Median Est. Debt-to-GDP	(0.099)	(0.098)	(0.099)	(0.099)	(0.098)
Treatment	0.117	0.017	0.120*	0.073	0.112
	(0.075)	(0.074)	(0.071)	(0.071)	(0.072)
Observations	1612	1612	1612	1611	1610
Panel D: Underestimators					
Treatment ×	-0.284	-0.007	0.342*	-0.073	0.194
Underestimate	(0.200)	(0.175)	(0.190)	(0.184)	(0.190)
Treatment	0.310	0.033	-0.266	0.050	-0.085
	(0.193)	(0.167)	(0.182)	(0.176)	(0.183)
Observations	1612	1612	1612	1611	1610

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Each Panel shows results for a different dimension of heterogeneity. All estimations are based on the pooled data from the representative sample and from MTurk. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent, the respondent's number of children (top-coded at five), and a dummy for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.27: Beliefs about political bias

	Left-wing Biased	Right-wing Biased	No political Bias
Panel A: Pooled			
Treatment	0.001	0.013	-0.014
	(0.011)	(0.014)	(0.017)
Adjusted p-value	[1.000]	[1.000]	[1.000]
Observations	1615	1615	1615
Panel B: Rep. Sample			
Treatment	-0.017	0.020	-0.003
	(0.013)	(0.016)	(0.021)
Adjusted p-value	[0.486]	[0.486]	[0.486]
Observations	813	813	813
Panel C: MTurk			
Treatment	0.018	0.009	-0.027
	(0.017)	(0.023)	(0.028)
Adjusted p-value	[1.000]	[1.000]	[1.000]
Observations	802	802	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Panel A shows estimations on the pooled sample, Panel B shows estimations on the representative sample and Panel C shows estimations on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses. False-discovery rate adjusted p-values are in brackets.

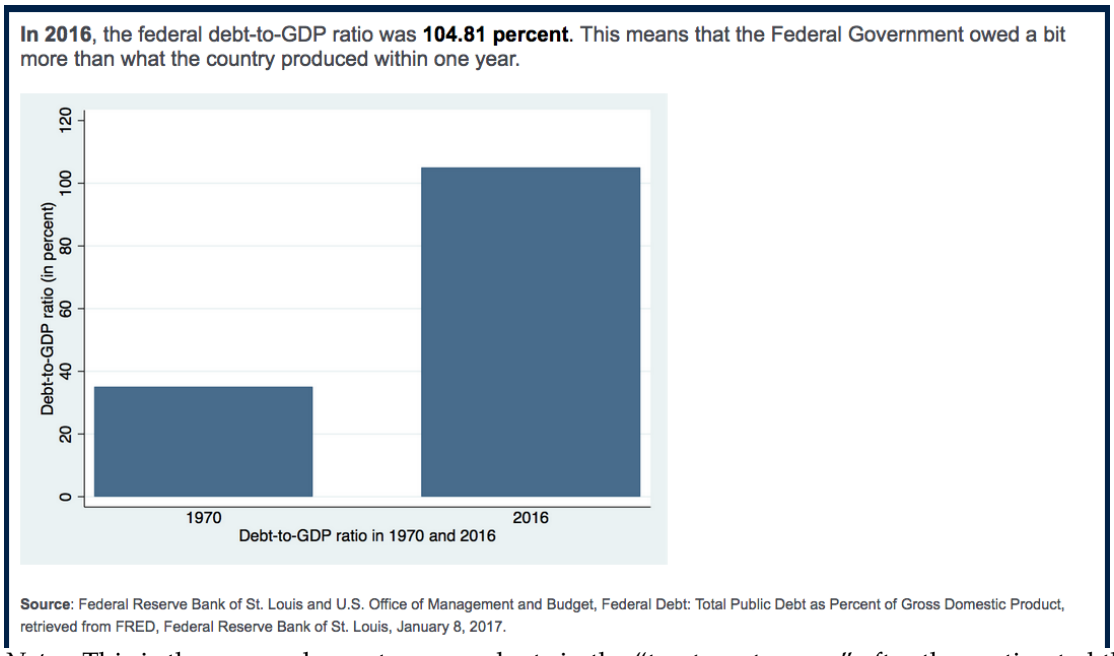
Table A.28: Correlates of beliefs about the debt-to-GDP ratio

	Perc. Debt-to-GDP		
	Pooled Sample	Rep. Sample	MTurk Sample
Male	5.904***	9.373***	3.681
	(1.832)	(2.447)	(2.812)
Age	0.234***	0.142	0.323**
	(0.084)	(0.103)	(0.142)
Log(Income)	1.489	0.751	2.241
	(1.125)	(1.448)	(1.795)
Number of children	-0.978	-1.066	-1.089
	(0.830)	(0.994)	(1.456)
Employed Full-Time	2.732	-1.545	8.779
	(3.340)	(4.082)	(5.867)
Employed Part-Time	8.210**	8.298	11.682*
	(4.128)	(5.561)	(6.715)
Unemployed	0.779	-0.380	3.876
	(4.302)	(5.944)	(6.733)
Retired	0.278	-2.570	7.095
	(4.284)	(4.946)	(10.757)
Student	1.880	-5.477	9.597
	(4.991)	(6.453)	(8.023)
High Education	0.500	-0.617	0.954
	(1.893)	(2.548)	(2.803)
Republican	2.909	3.247	2.291
	(1.946)	(2.505)	(3.137)
Observations	1615	813	802

Notes: Column (1) shows the estimation on the pooled sample, column (2) shows the estimation on the representative sample and column (3) shows the estimation on the MTurk sample. All specifications control for age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

3.C Additional Figures

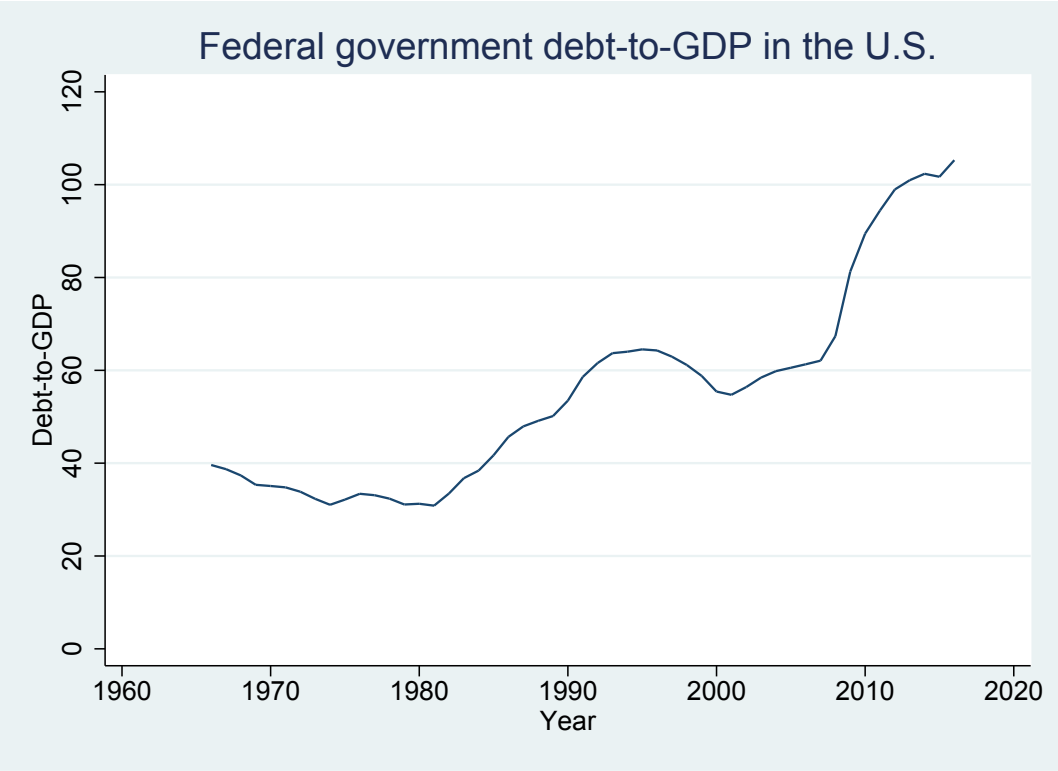
Figure A.1: Treatment screen



Notes: This is the screen shown to respondents in the "treatment group" after they estimated the debt-to-GDP ratio.

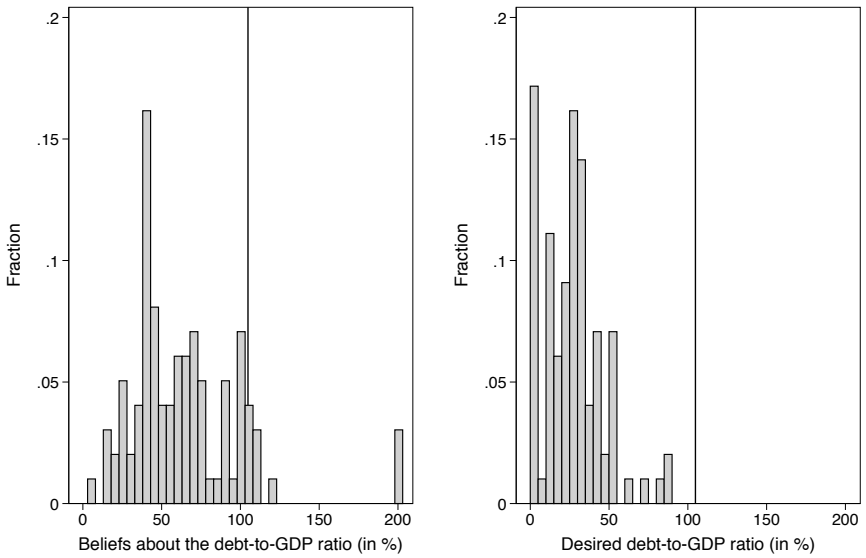
3.C.1 Beliefs about the debt-to-GDP ratio

Figure A.2: Evolution of debt-to-GDP ratio



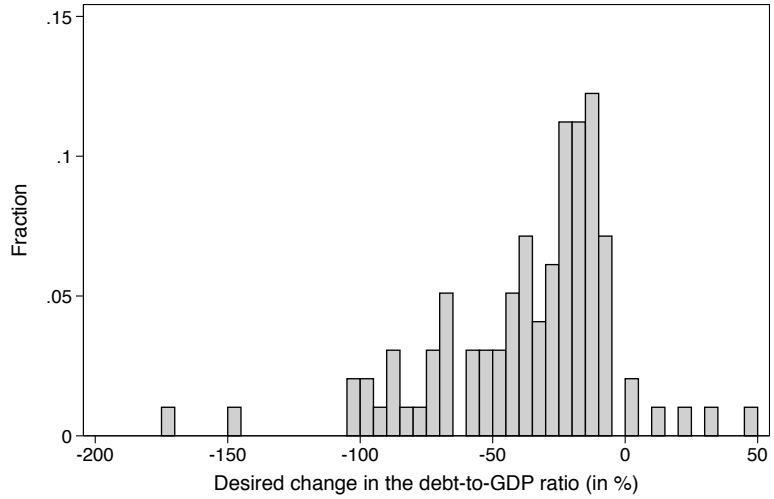
Notes: In this figure we display the evolution of the federal government debt-to-GDP ratio in the US from 1965 until 2016. Source: FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GFDEGDQ188S, July 24, 2017..

Figure A.3: Beliefs about the debt-to-GDP ratio and desired debt-to-GDP ratio (pilot experiment)



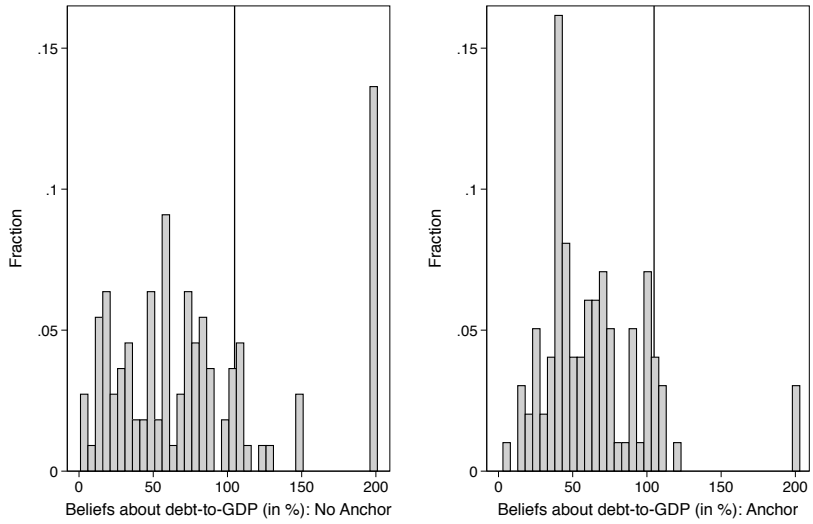
Notes: This figure describes the distribution of beliefs about the debt-to-GDP ratio in a pilot experiment with 200 respondents that we ran on MTurk. The figure is based on data from half of the respondents which we provided with a historical anchor. On the left we display people's beliefs about the debt-to-GDP ratio. On the right we show people's desired debt-to-GDP ratio. The estimated and desired debt-to-GDP ratios are winsorised at 200 percent. The median estimate is 56.23 percent and the median desired debt-to-GDP ratio is 25 percent.

Figure A.4: Desired change in debt-to-GDP ratio (pilot experiment)



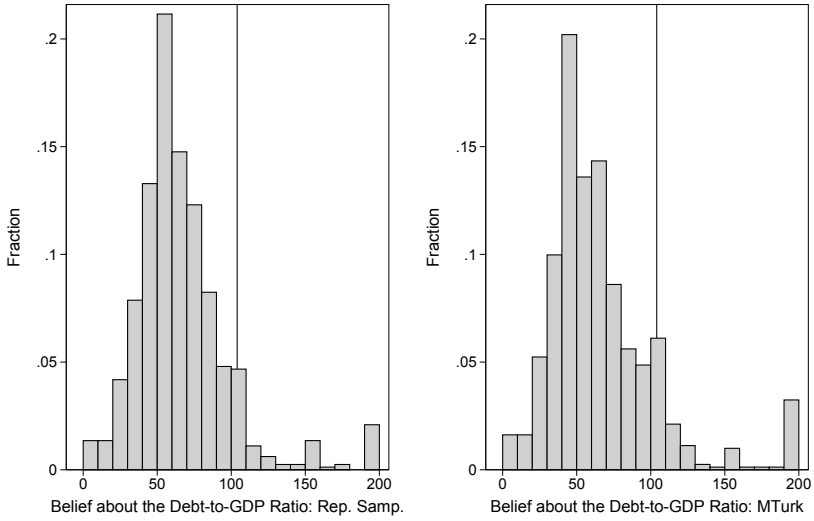
Notes: This figure describes the distribution of desired changes in the debt-to-GDP ratio in a pilot experiment with 200 respondents that we ran on MTurk. The figure is based on data from half of the respondents which we provided with a historical anchor.

Figure A.5: Beliefs about the debt-to-GDP ratio with and without anchor (pilot experiment)



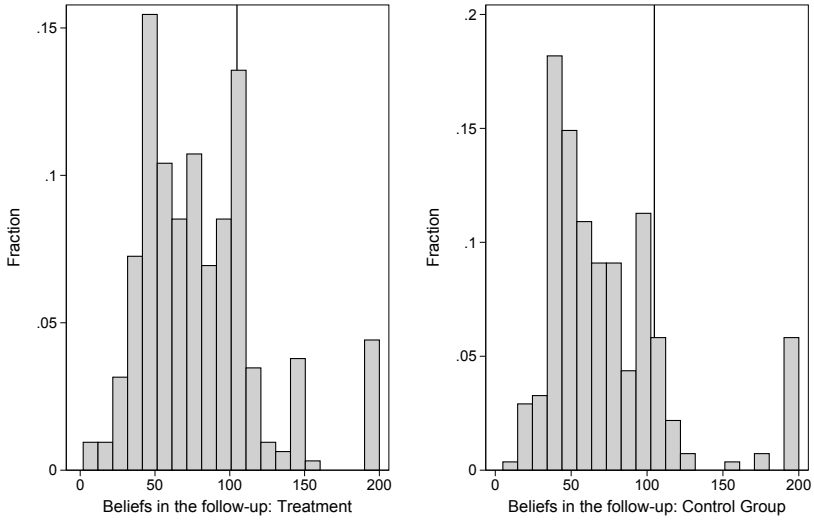
Notes: This figure describes the distribution of beliefs about the debt-to-GDP ratio in a pilot experiment with 200 respondents that we ran on MTurk. On the left we display people's beliefs when they are not given an anchor. On the right we display their beliefs when they are given a historical anchor. The estimates are winsorised at a debt-to-GDP ratio of 200 percent. The median estimate is 61.5 percent (56.23 percent) without (with) historical anchor.

Figure A.6: Beliefs about the debt-to-GDP ratio (representative sample and MTurk)



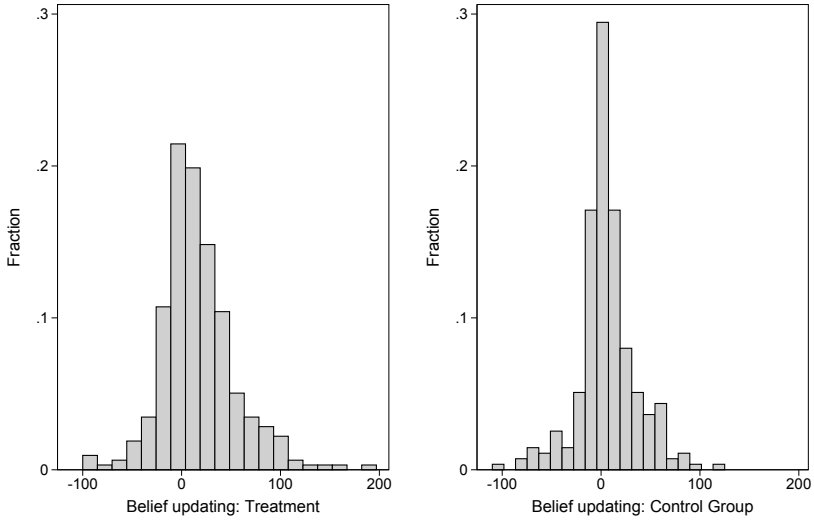
Notes: In this figure we display people's beliefs about the current debt-to-GDP ratio separately for the representative sample and for the MTurk sample. The estimates are winsorised at a debt-to-GDP ratio of 200 percent.

Figure A.7: Beliefs about the debt-to-GDP ratio in the follow-up



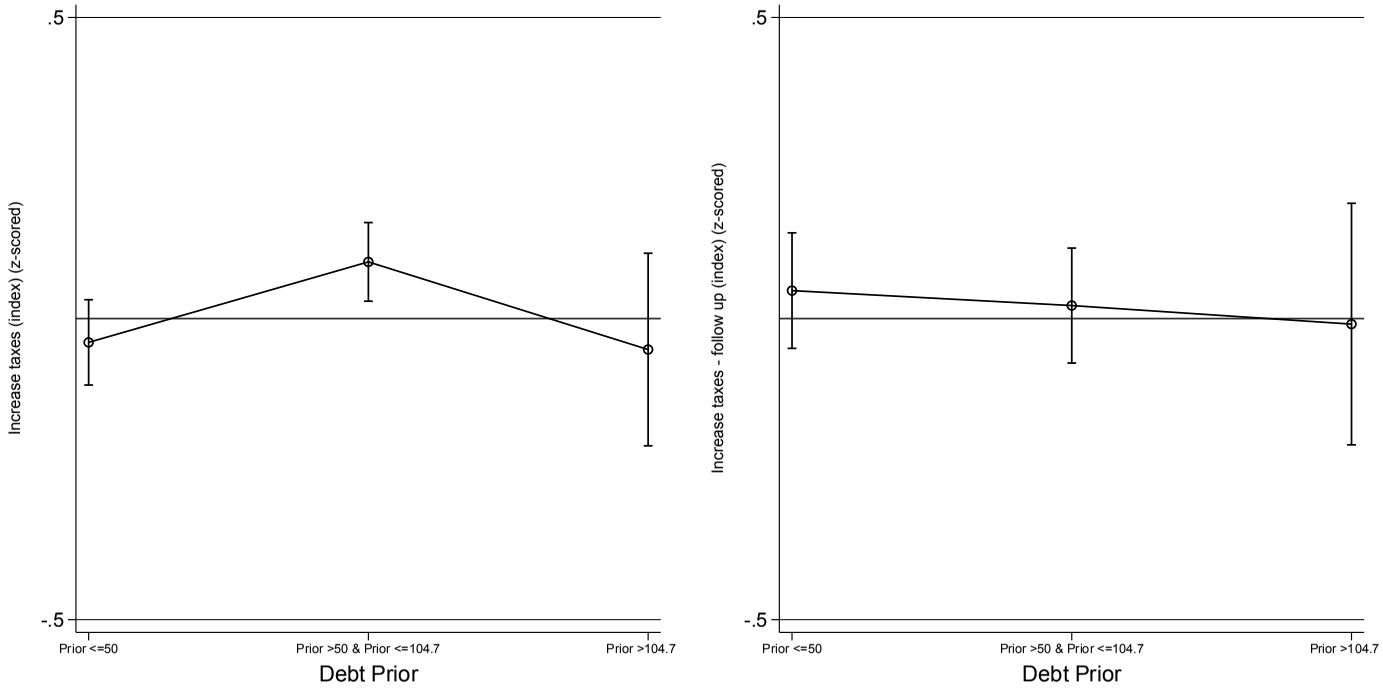
Notes: This figure describes the distribution of beliefs about the debt-to-GDP ratio in the four-week follow-up experiment. The estimates are winsorised at a debt-to-GDP ratio of 200 percent.

Figure A.8: Belief updating



Notes: This figure describes the distribution of changes in beliefs about the debt-to-GDP ratio between the main experiment and the four-week follow-up experiment across the two treatment arms.

Figure A.9: Heterogeneous effects on attitudes towards taxation (index): by prior beliefs



Notes: This figure describes treatment effects on an index of attitudes towards taxation by people's prior beliefs about the debt-to-GDP ratio. The outcome variable is z-scored using the mean and standard deviation in the control group. The figure displays the point estimate of the treatment effects with 90 percent confidence intervals. The figure on the left is based on pooled data from the main experiments on the representative sample and MTurk, while the figure on the right is based on the follow-up survey on MTurk. The treatment effect estimates control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample.

3.D Additional results

3.D.1 Effects of information on support for tax- vs. debt-financed spending program

We also ask our respondents about their support for an infrastructure program and randomly assign whether this program is tax-financed or debt-financed. This allows us to identify whether people's beliefs about government debt affect their support for new spending programs depending on the proposed mode of financing.

To analyze whether our information treatment has differential effects depending on whether a proposed spending program is tax-financed or debt-financed, we create the dummy variable $Debt_i$, which takes value one for participants who are asked about support for a debt-financed infrastructure program and value zero for participants who are asked about support for a tax-financed program. We estimate the following specification:

$$y_i = \pi_0 + \pi_1 Treatment_i \times Debt_i + \pi_2 Treatment_i + \pi_3 Debt_i + \Pi^T \mathbf{X_i} + \varepsilon_i$$

The coefficients π_1 and π_2 capture effects of our treatment on support for the program that potentially differ depending on the mode of financing. The coefficient π_3 captures whether people in the control group differentially support tax-financed and debt-financed infrastructure investments.

Table A.29 shows that people are more likely to support a government in-

frastructure investment program if it is financed by a temporary tax increase rather than by issuing new debt. We find no evidence that learning about the actual debt-to-GDP ratio affects people's support for this investment program – irrespective of the mode of financing.

This result differs from our previous finding that learning about the debt-to-GDP ratio decreases people's demand for government spending. We believe that this could be the case as (i) we have less variation available in the measure of people's support for the infrastructure program, (ii) we have less statistical power and (iii) the framing of the question on infrastructure spending is different.

Table A.29: Debt vs tax-based financing of an infrastructure investment program

	Support Infr. Inv. Program		
	Pooled Sample	Rep. Sample	MTurk Sample
Debt- vs. Tax-based Financing			
Debt-financed × Treatment	-0.012	0.035	-0.084
	(0.097)	(0.139)	(0.137)
Treatment	0.055	0.118	0.014
	(0.069)	(0.101)	(0.097)
Debt-financed	-0.265***	-0.341***	-0.184*
	(0.069)	(0.096)	(0.099)
Observations	1611	809	802

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Column (1) shows the estimation on the pooled sample, column (2) shows the estimation on the representative sample and column (3) shows the estimation on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Table A.30: Debt vs. tax-based financing of an infrastructure investment program: Heterogeneity in control group

	Support Infr. Inv. Program		
	Pooled Sample	Rep. Sample	MTurk Sample
Debt- vs. Tax-based Financing			
Debt-financed × Debt-to-GDP	-0.394**	-0.451*	-0.406
	(0.188)	(0.267)	(0.255)
Debt-financed × Republican	-0.074	-0.038	-0.000
	(0.148)	(0.198)	(0.223)
Debt-financed × High Education	-0.069	0.011	-0.113
	(0.145)	(0.207)	(0.206)
Debt-financed × High Income	-0.026	-0.148	0.055
	(0.147)	(0.206)	(0.210)
Debt-financed × Age	0.002	0.004	0.010
	(0.004)	(0.006)	(0.009)
Debt-financed × Children	-0.057	0.144	-0.335
	(0.146)	(0.203)	(0.212)
Debt-financed	-0.374*	-0.652*	-0.413
	(0.214)	(0.335)	(0.322)
Debt-to-GDP	0.134	0.221	0.021
	(0.141)	(0.209)	(0.170)
Republican	-0.260**	-0.085	-0.531***
	(0.105)	(0.142)	(0.156)
High Education	0.066	-0.048	0.149
	(0.105)	(0.151)	(0.145)
High Income	0.177	0.156	0.268
	(0.136)	(0.192)	(0.198)
Age	-0.003	-0.004	-0.005
	(0.004)	(0.005)	(0.006)
Children	0.028	0.026	0.079
	(0.105)	(0.148)	(0.155)
Observations	811	427	384

Notes: The outcome variables are z-scored using the mean and standard deviation in the control group. Column (1) shows the estimation on the pooled sample, column (2) shows the estimation on the representative sample and column (3) shows the estimation on the MTurk sample. All specifications control for the perceived debt-to-GDP ratio (winsorised at 200 percent), age, gender, a dummy for whether the respondent has at least a bachelor degree, the log of total household income, dummies for full-time employment, part-time employment, unemployment, retirement, full-time education and other employment status, the respondent's trust in official US government statistics, dummies for being a Republican or an Independent as well as the respondent's number of children (top-coded at five). The estimations on the pooled sample also control for whether the respondent is part of the representative sample or the MTurk sample. Robust standard errors are in parentheses.

Chapter 4

Conspicuous Consumption and Peer Effects: Evidence From a Randomised Field Experiment

Christopher Roth[1]

Abstract How do changes in our peers' consumption choices affect our consumer behavior? Combining novel Indonesian data on the visibility of consumption goods with a randomised cash transfer program, I find that the expenditure share of visible goods rises for households whose peers' income increases. The results are inconsistent with a signalling model, but consistent with a model of

[1] I would like to thank Johannes Abeler, Robert Akerlof, Abhijit Banerjee, Seth Blumberg, Sam Bowles, Stefano Caria, Rachel Cassidy, Damian Clarke, Ian Crawford, Giacomo De Giorgi, Marcel Fafchamps, James Fenske, Alexis Grigorieff, Ori Heffetz, David Huffman, Clément Imbert, Michael Koelle, Omer Moav, Ricardo Perez-Truglia, Simon Quinn, Climent Quintana-Domeque, Gautam Rao, Imran Rasul, Russel Toth and many others for help. I would like to thank seminar audiences at Barcelona GSE, CSAE, NEUDC, PacDev and the RES for useful comments. I am particularly grateful to Vivi Alatas, Matt Wai-Poi and the World Bank for providing me with the PKH conditional cash transfer dataset. For the data collection on the consumption visibility data I received approval from the ethics committee of the University of Oxford.

relative concerns over visible goods. Finally, I shed light on the social mechanisms underlying peer effects in consumption.

4.1 Introduction

People compare themselves with others on the basis of their social and economic status (Card et al., 2012; Frank, 1985; Luttmer, 2005; Mas and Moretti, 2009; Perez-Truglia, 2016). Such social comparisons have implications for people's consumption behaviour as people may use conspicuous consumption as a way to signal their economic status to others (Bagwell and Bernheim, 1996; Bursztyn et al., 2018; Glazer and Konrad, 1996; Veblen, 1899).

Both a theoretical literature and an empirical literature have highlighted the importance of understanding conspicuous consumption to explain important economic phenomena, such as poverty traps (Moav and Neeman, 2012), indebtedness and consequent bankruptcy (Agarwal et al., 2016) or wealth inequality between Blacks and Whites in the United States (Charles et al., 2009).

However, little direct evidence exists on whether people adjust their expenditure share of visible goods in response to exogenous income and consumption increases of their peers. This paper uses data from a randomised conditional cash transfer program for poor households from Indonesia to test whether people change the share of visible consumption in response to an increase in peer group income.

I exploit the partial population design of an Indonesian cash transfer program, in which a subset of the population in a given village receives an

increase in income for a duration of six years. I compare untreated households in treated villages with comparable households from untreated villages. This allows me to study how consumption patterns of households are affected by an exogenous increase in reference group income.[2]

To assess consumption visibility, I use two survey instruments used in the literature (Charles et al., 2009; Heffetz, 2011). I collect this consumption visibility data to obtain a ranking of consumption visibility in Indonesia. This in turn allows me to analyze the consumption patterns of households in terms of visible expenditures (e.g. clothing) and nonvisible expenditures (e.g. food at home).

I find that a rise in reference group income increases a household's visible expenditure share, but not their total expenditures. Thus, households reallocate some of their expenditures from nonvisible to visible goods. Visible expenditure shares increase by two percentage points or about 20 percent, while nonvisible expenditure shares decrease significantly. The results are consistent with a simple model of relative concerns over status goods, but insonsistent with a standard signalling model of conspicuous consumption, which predicts that the share of visible consumption decreases when peer group income increases (Charles et al., 2009).

Subsequently, I examine heterogeneity in treatment effects. As the key assumption underlying peer effects in conspicuous consumption is the relevance of social image concerns, I provide evidence on the social mechanisms underlying peer effects in consumption. The increases in visible expenditures of untreated households in treated villages are larger in villages with lower levels of social activities, consistent with the claim by Veblen (1899)

[2]This enables me to overcome the reflection problem (Manski, 1993), i.e. the challenge of separately identifying genuine peer effects from correlated effects.

that conspicuous consumption is higher in communities with lower social capital and more anonymous interactions.

Then, I provide evidence that the results are unlikely driven by program recipients sharing cash with untreated households in their village (Angelucci and De Giorgi, 2009). Specifically, I show that there is no significant treatment heterogeneity by a proxy for sharing norms. Finally, I show that the treatment effects are not driven by supply-side effects, local prices or non-linear Engel curves

This paper contributes to several literatures. First, it contributes to the literature on neighbourhood effects (Kling et al., 2007), social image concerns (Bursztyn and Jensen, 2015; DellaVigna et al., 2017; Perez-Truglia and Cruces, 2017), peer effects (Bandiera et al., 2009; Bertrand et al., 2000; Bursztyn et al., 2014; Dahl et al., 2014; Duflo and Saez, 2003; Mas and Moretti, 2009; Sacerdote, 2001) and, in particular, peer effects in consumption (Bertrand and Morse, 2016; De Giorgi et al., 2016; Grinblatt et al., 2008; Kuhn et al., 2011; Moretti, 2011).

My paper is most closely related to Kuhn et al. (2011) who use the "Dutch Postcode lottery" as a source of exogenous variation in neighbours' income. Their seminal paper is the first natural experiment on peer effects in consumption to address the reflection problem. In their paper, lottery winners win a cash prize and one winner in a given postcode also receives a BMW. Kuhn et al. (2011) show that neighbours of recent lottery winners by and large do not change their consumption levels except for increases in their spending on cars. They do not find robust increases in other possibly conspicuous goods such as expenditures on food out, exterior renovations and expenditures on vacations and they do not systematically investigate

whether consumption shares of visible goods increase. My study is different from Kuhn et al. (2011) in many important respects: My paper constitutes a more direct test of peer effects in visible consumption, as I investigate how consumption shares of visible goods are affected by an increase in peer income. Moreover, my paper sheds light on the social mechanisms underlying peer effects, by exploiting rich data on social activities. Finally, my data is from a very different context: while Kuhn et al. (2011) studies household behaviour in the Netherlands, I focus on poor household from rural Indonesia.

My paper is also related to Bertrand and Morse (2016) who show that the consumption of non-rich households significantly correlates with the income of the top 10 percent of earners in the households' state. The share of visible goods of non-rich households increases in the income of the top earners, consistent with the field-experimental evidence from my paper.

Second, this paper contributes to the literature on indirect effects of government programs on the consumption behaviour of program ineligibles (Angelucci and De Giorgi, 2009). My study is similar to Angelucci and De Giorgi (2009), who use a conditional cash transfer program from Mexico. They show that the overall food consumption of program ineligibles in treated villages increases. Angelucci and De Giorgi (2009) argue that the most likely explanation for the patterns in the data is that eligible households shared the cash transfer with ineligible households. Their findings are in contrast to the results from my paper: While Angelucci and De Giorgi (2009) document potentially welfare-enhancing spillovers of the program through sharing, my paper emphasizes potentially welfare-reducing spillovers as a result of status concerns.

Third, my paper contributes to the literature on visibility of consumption (Heffetz, 2011; Jakiela and Ozier, 2016). Few studies have systematically exploited consumption visibility and status concerns (Bertrand and Morse, 2016; Charles et al., 2009; Kaus, 2013). Bursztyn et al. (2018) identifies people's demand for status goods using field experiments in Indonesia. Heffetz (2011) shows that consumption visibility explains up to one third of the income elasiticities of various consumption goods. Charles et al. (2009) provide evidence on racial differences in conspicuous consumption. They show that the share of visible goods is negatively correlated with mean reference group income, consistent with a simple model of signalling.

This paper is structured as follows. Section 4.2 outlines a signalling model of conspicuous consumption and a model of relative concerns over visible goods. Section 4.3 presents the data. Then in Section 4.4, I present the main results. Subsequently, Section 4.5 examines social mechanisms and assesses alternative explanations of results. Section 4.6 examines robustness. Section 4.7 concludes.

4.2 Conceptual framework

I outline the predictions of two models that could explain peer effects in conspicuous consumption. First, I outline the predictions of a signalling model of conspicuous consumption (Charles et al., 2009; Ireland, 1994; Perez-Truglia, 2013), which relies on the idea that individuals may use visible consumption as a way to signal their unobserved income (Heffetz, 2011; Heffetz and Frank, 2008). Second, I outline the predictions of a simple model in which people have direct concerns over their relative consumption of visible goods.

4.2.1 A signalling model

In this section I outline a signalling model of conspiuous consumption which closely follows Charles et al. (2009) and Perez-Truglia (2013).[3] The signalling theory of conspicuous consumption posits that people use visible consumption in order to signal their unobserved income to others. The theory posits that these signals of income are helpful to obtain social status, s, which may be useful to obtain goods that are not traded in markets, such as respect, admiration or social relationships. First, we need to characterize people's reference group and the income distribution:

Assumption 1 (REFERENCE GROUP AND INCOME HETEROGENEITY)**.** *In my setting a reference group is comprised by a continuum of individuals from the same village. There is heterogeneity in income, which is distributed according to a non-degenerate cumulative function $F_y(.)$ over the support $[y_0, y_1]$ with mean income given by μ_g. An individual can spend its income on either visible goods, v, or nonvisible goods, $y - v$.*

Next, we define people's preferences over goods:

Assumption 2 (PREFERENCES OVER GOODS)**.** *A person's preferences over these goods are given by*

$$U_i = u(y_i - v_i) + \alpha u(v_i) + s \tag{4.1}$$

where u is an increasing and concave function; v_i is visible consumption; $y_i - v_i$ is non-visible consumption; and s is social status.

[3]For more elaborate treatments of signalling models of conspicuous consumption see Glazer and Konrad (1996) and Ireland (1994).

Remark We focus on people's signalling of unobserved income which entails status. A natural assumption is that people's unobserved income is evaluated against the average income in the individual's group. People who want to make an inference about an individual's unobserved income observe both the individual's visible consumption, v_i, and the average group income, μ_g. In equilibirum people's inferences about people's incomes are correct. In other words, specific levels of visible consumption correspond to a particular income. An individual's status is given by: $s_i = (v_i, \mu_g)$

Let $v^*(y_i, \mu_g)$ be an individual's equilibrium level of visible consumption, where y_i denotes an individual's income. Differentiating equation 4.1 with respect to visible consumption, subject to a budget set yields the first order condition:

Proposition 1 (FIRST-ORDER CONDITION)**.** *The first oder condition is given by:*

$$\frac{\partial s(v^*, \mu_g)}{\partial v^*} = u'(y_i - v^*) - \alpha u'(v^*) \tag{4.2}$$

Remark The above equation has two key implications. First, people's conspicuous consumption is increasing in people's own income. Second, people's level of visible consumption is higher when they have status concerns.

Comparative statics How does people's visible consumption depend on group income μ_g?

Proposition 2 (PEER INCOME)**.** *Differentiating both sides of the first order condition with respect to μ_g yields:*

$$\frac{\partial^2 s(v^*, \mu_g)}{\partial v^* \partial \mu_g} = \frac{-\partial v^*(y_i, \mu_g)}{\mu_g}[u''(y_i - v^*) + \alpha u''(v^*) + \frac{\partial^2 s}{\partial v^* \partial v^*}] \qquad (4.3)$$

The above equation implies that the effect of group income on visible consumption, $\frac{\partial v^*(y_i,\mu_g)}{\mu_g}$, is of the same sign as the second derivative of status $\frac{\partial^2 s(v^*,\mu_g)}{\partial v^* \partial \mu_g}$. People's visible consumption decreases in their group income if $\frac{\partial^2 s(v^*,\mu_g)}{\partial v^* \partial \mu_g} < 0$.

Remark The effect of group income on status is dependent on the exact way status is allocated in society. Under simple and standard assumptions in the happiness literature, an individual's status, s_i, decreases in her group income, μ_g. If $\frac{\partial s_i}{\partial \mu_g} < 0$, then visible consumption decreases in average group income. The intuition for this result is that the higher an individual's reference group income, the less an individual gains compared to the mean from an increase in visible expenditure.[4]

4.2.2 A model of relative concerns over visible consumption

We now focus on a model where individuals have relative and absolute concerns for visible goods and only absolute concerns for nonvisible goods.

Assumption 3 (PLAYER UTILITY)**.** *There are two identical players*[5]*, player i and player j, who each noncooperatively maximize their utility function subject to their*

[4]While the assumption outlined above is reasonable and has been made in influential papers on conspicuous consumtion (e.g. Charles et al. (2009)), it is also conceivable that belonging to a wealthier reference group could convey higher status, i.e. $\frac{\partial s_i}{\partial \mu_g} > 0$. For an example, see Ireland (1994).

[5]By players I mean households which shall be conceived of as a single decision unit.

budget constraint:

$$U_i = \theta \ln(n_i) + (1-\theta)\ln(v_i + \alpha(v_i - v_j)), \textit{where } \alpha > 0 \;\; \textit{s.t.} \quad n_i + v_i p_v = y_i. \tag{4.4}$$

Remark Both players care about nonvisible goods only in absolute terms, whereas they care about visible goods both in absolute and relative terms. The price of nonvisible goods is normalised to 1, the price of visible goods is given by p_v, and an individual's income is given by y_i. Individual j's utility function and budget constraint are identical to i's utility function and budget constraint. As the relative concerns parameter is positive, $\alpha > 0$, an increase in player $j's$ level of visible consumption decreases player i's utility and vice versa. Further, I assume that individuals care about both visible and nonvisible goods, i.e. $\theta \in (0, 1)$. Crucially, I make the following two restrictions on the players' levels of visible consumption to ensure non-negativity of the argument of the log:

$$(1+\alpha)v_i \geq \alpha v_j; \tag{4.5}$$

$$(1+\alpha)v_j \geq \alpha v_i. \tag{4.6}$$

As will be shown subsequently, these conditions are satisfied in equilibrium.

Assumption 4 (INFORMATION STRUCTURE)**.** *The players' rationality and their preferences are all common knowledge.*

The solution concept used is Nash equilibrium. Solving the model's constrained optimisation problem yields the following best response function for player i:

Proposition 3 (BEST RESPONSE FUNCTIONS)**.** *Player i's best response is given*

by:

$$v_i(v_j) = \left(\frac{\theta\alpha}{1+\alpha}\right) v_j + (y_i/p_v)(1-\theta). \tag{4.7}$$

Remark Individual i's level of visible consumption is increasing in $j's$ level of visible consumption, i.e. v_i and v_j are strategic complements. The effect of an increase in j's level of visible consumption can be expressed succinctly as follows:

$$\frac{\partial v_i}{\partial v_j} = \frac{\theta\alpha}{1+\alpha} > 0 \text{ given } \alpha, \theta > 0. \tag{4.8}$$

The intuition for this result is as follows: an increase in j's visible consumption increases i's marginal utility from consuming v_i. Therefore, v_i is increasing in v_j and vice versa. The magnitude of effects will depend positively on the magnitude of relative concerns, α, as well as the relative weight θ put on nonvisible consumption.

For equilibrium characterisation, I assume that $y_i = y_j = y$ to simplify the calculations. Exploiting the resulting symmetry of the problem, I solve for the unique Nash equilibrium value of visible consumption given by:

$$v_i^* = v_j^* = \left(\frac{(1+\alpha)(1-\theta)}{1+\alpha(1-\theta)}\right)(y/p_v). \tag{4.9}$$

Crucially, the non-negativity constraints are satisfied as $v_i^* = v_j^*$. Plugging the equilibrium value into the budget constraint yields the equilibrium value of nonvisible consumption:

$$n_i^* = n_j^* = \frac{\theta y}{1+\alpha(1-\theta)}. \tag{4.10}$$

Comparative statics The above equations imply that both i's visible and i's nonvisible consumption are increasing in i's income; Moreover, i's visible consumption increases in i's relative concerns α.

Proposition 4 (EFFECT OF AN INCREASE IN j'S INCOME ON i'S CONSUMPTION)**.** *The visible consumption of i is increasing in j's income and vice versa, whereas i's nonvisible consumption is decreasing in j's income:*

$$\frac{\partial v_i(v_j(y_j))}{\partial y_j} = \frac{\partial v_i}{\partial v_j}\frac{\partial v_j}{\partial y_j} = \left(\frac{\theta\alpha}{1+\alpha}\right)(1/p_v)(1-\theta) > 0; \quad (4.11)$$

$$\frac{\partial n_i(v_i(v_j(y_j)))}{y_j} = \frac{\partial n_i}{\partial v_i}\frac{\partial v_i}{\partial v_j}\frac{\partial v_j}{\partial y_j} = -\left(\frac{\theta\alpha}{1+\alpha}\right)(1-\theta) < 0. \quad (4.12)$$

Remark This model thus predicts that own visible consumption and reference group visible consumption are strategic complements as long as $\alpha > 0$. Many models of social learning would also predict that own visible consumption increases in reference group visible consumption. In these models individuals learn about the product quality and they make inferences about the product quality based on their peers' consumer behaviour (Bikhchandani et al., 1992).

4.2.3 Hypotheses

The model of relative concerns over status goods and models of social learning predict that people should increase their spending on visible goods in response to an increase in peer group income. The signalling model, on the other hand, predicts that people's visible consumption share decreases in peer group income.

4.3 Data

Background I use the Program Keluarga Harapan (PKH) conditional cash transfer program dataset that was collected in Indonesia in 2007 and 2009 (Alatas, 2011). In 2007 the PKH pilot was started in five provinces: West Java, East Java, North Sulawesi, Gorontalo as well as East Nusa Tenggara (NTT). PKH delivers cash transfers to poor households for six years. The size of the transfers is substantial: households received amounts between 60 and 220 US dollars per year. Average yearly payments to households are 130 US Dollars, or about 12 percent of the households' yearly expenditure in the year before the PKH program started.

Survey data The conditional cash transfer dataset is unusually rich: the data ranges from socio-economic and demographic characteristics such as schooling and health to labour market outcomes and data on various consumption goods.[6] In particular, we use monthly consumption data for all goods except for food expenditures, which are weekly.[7] Total monthly household expenditures are defined as the sum of all available household level monthly expenditures which will be described later. Consumption shares are defined as the monthly consumption expenditures divided by total monthly expenditures. The baseline survey was fielded between June and August 2007; the follow-up survey was conducted between October and December 2009. The attrition rate is 3.4 percent and does not differ across treatment arms.

[6]Data was collected by an independent institution, the University of Gadjah Mada.

[7]We then calculate monthly food expenditures by multiplying weekly food expenditures by 4.3.

4.3.1 Experimental design

The study is based on a partial-population design, i.e. only a subset of the population receives the treatment. Specifically, there are completely untreated villages as well as treated and untreated households in treated villages. This paper focuses on the rural subsample of the PKH program. Since the focus of the program is poverty alleviation, upper income quintile districts were excluded from PKH eligibility based on poverty rates, malnutrition and schooling records. Out of eligible regions only those with sufficient health and education service institutions were chosen. Then out of the list of 85 eligible rural sub-districts 50 were randomly selected for treatment and 35 were chosen to be the control group. Political pressures and an unexpected expansion of the program on East Java resulted in deviations of the actual allocation of the PKH program from the original allocation. 8 out of the 35 sub-districts that were designated to be part of the control group received PKH funds. This may result in bias as the unobservables within contaminated sub-districts that induced a departure from the initial implementation plan, simultaneously affected household responses. We deal with this by using the original treatment assignment as an instrument for the actual treatment assignment.

Baseline Balance In a first step, this section tests the implications of the randomisation for the demographic characteristics, consumption levels and asset holdings of households in treated and untreated sub-districts at baseline. First, I consider whether baseline balance holds for the original assignment and then in a second step whether it holds for the actual assignment. I do so comparing the means and allowing the standard errors to be clus-

tered at the sub-district level. The variables of interest include the total, visible, nonvisible consumption expenditure as well as the shares spent on visible and nonvisible goods; demographic characteristics, such as age, education, employment sector, household size, social activities; and village level characteristics, such as the percentage of households without electricity, percentage of farmers, population size and the prevalence of social punishment norms. I find almost no statistical differences between households in sub-districts originally and actually assigned to be treated vs. households in sub-districts originally and actually assigned to be untreated respectively.[8] Thus, the randomisation seems to have worked well and the treatment contamination does not seem to have systematically resulted in imbalances. Importantly all outcome variables of interest, i.e. consumption levels and shares are balanced at baseline. The baseline balance of treated and untreated sub-districts is evidenced in Table 4.1. Balance induced by the original treatment assignment is shown in Table A.1 in Appendix B. Moreover, as can be seen in Table A.8 in Appendix B, I find barely any statistical differences between contaminated and uncontaminated sub-districts.[9]

Selection process at the household level At the sub-district level, the program was offered to households that satisfied both certain demographic as well as certain economic requirements. The initial list of households considered for the program was based on a list from the unconditional cash transfer program from 2005. Approximately 55 percent of people from this initial list were included in the list of potential PKH beneficiaries. Using demographic data, households which met the demographic requirements

[8] In the original assignment the share of those with previous government aid is higher.

[9] I find that the share of those without schooling and of those working in the agricultural sector is lower in the contaminated sub districts.

of the PKH program (households with pregnant and/or lactating women; households with children aged 0-15 years; and households with children aged 16-18 years who have not yet completed 9 years of basic education) were selected. Then, standard demographics and wealth indicators (e.g. assets) were used in the proxy-means models which determined eligibility status (Alatas, 2011). In the end the program covered 11 percent of the households per subdistrict.

4.3.2 New consumption visibility data

I collected consumption visibility data in Indonesia based on the surveys used by Heffetz (2011) and Charles et al. (2009) in August and September 2013. I did this for two reasons: first, I had to define some consumption categories as "visible" and others as "non-visible", and I wanted to inform this from the local context. Second, I wanted to investigate whether consumption visibility in a developing country is different from consumption visibility in the western world.

To the best of my knowledge, I am the first to collect consumption visibility data of this kind in a developing country. Furthermore, this is the first time that the same individuals responded to both visibility surveys used in the literature. Thus, this is the first attempt to validate both visibility survey instruments from the literature (Charles et al., 2009; Heffetz, 2011). The surveys were translated into Bahasa Indonesia and the consumption categories were adapted to the cultural idiosyncrasies of Indonesia. In particular, I included spending on weddings as well as spending to go on pilgrimage as

Table 4.1: Baseline characteristics of households in treated and untreated sub-districts

	Treated sub-districts	Untreated sub-districts	P-value
Consumption			
Total Log expenditure	13.61	13.65	.295
Total expenditure	96.32	99.00	.591
Log Visible expenditure	11.16	11.23	.400
Visible expenditure	10.31	10.97	.397
Log Nonvisible expenditure	13.16	13.20	.341
Nonvisible expenditure	58.07	61.03	.293
Share Visible expenditure	0.11	0.11	.580
Share Nonvisible expenditure	0.67	0.67	.605
Average Demographics (Household Head)			
Main Profession: Agriculture	0.60	0.58	.585
Household Size	5.06	5.17	.463
Age of Household Head	44.59	44.58	.979
No Schooling	0.34	0.31	.497
Junior High School	0.10	0.10	.774
Senior High School	0.06	0.04	.195
Previous Government Aid	0.94	0.92	.168
Connection to Village Head	0.53	0.50	.566
No Sanction	0.65	0.69	.512
Average Village Characteristics			
Percentage Islam	92.37	95.11	.538
Population Size	4740	4632	.829
Percentage of Farmers	68.50	67.32	.689
Percentage without Electricity	0.02	0.02	.735
Ethnic Diversity (Herfindhal Index)	0.15	0.13	.606
Village Head Sanction	0.39	0.42	.629
N	1840	1480	3320

The expenditures in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar. Standard errors underlying the tests for the equality of means are clustered at the subdistrict level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.2: Baseline characteristics of treated and untreated households and controls

	Treated Subdistricts		**Control**	**P-Value Difference**	
	Eligibles	Ineligibles		A - C	B-C
Consumption					
	(A)	(B)	(C)		
Log Total expenditure	13.587	13.622	13.654	.151	.509
Log Visible expenditure	11.147	11.178	11.235	.302	.516
Log Nonvisible expenditure	13.161	13.167	13.205	.316	.393
Share Visible expenditure	0.113	0.107	0.112	.877	.342
Share Nonvisible expenditure	0.677	0.667	0.666	.342	.942
Demographics					
Main Profession: Agriculture	0.609	0.596	0.578	.528	.680
Household Size	5.120	5.021	5.163	.767	336
Age of Household Head	43.706	45.238	44.592	.184	.389
No Schooling	0.351	0.330	0.314	.388	.663
Junior High School	0.087	0.111	0.104	.283	.691
Senior High School	0.049	0.062	0.043	.663	.100*
Previous Government Aid	0.964	0.918	0.917	.002***	.912
Connection to Village Head	0.517	0.539	0.500	.754	.478
No Sanction	0.604	0.685	0.692	.211	.910
Observations	773	1067	1475	2248	2542

The expenditures in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar. Standard errors underlying the tests for the equality of means are clustered at the subdistrict level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

additional categories in the consumption visibility survey (for an overview see Appendix 4.B). I collected a convenience sample of 115 observations by visiting people in their homes as well as by asking people in public places in greater Jakarta and in villages around Yogyakarta. The mean duration of interviews was 35 minutes.[10] Efforts were made to make the survey as representative as possible by interviewing individuals from different income classes, with different jobs, different gender and ethnicities. I use Heffetz' measure of visibility which ranks expenditures by their consumption visibility. The question from Heffetz' survey was given as follows:

> *"Imagine that you meet a new person who lives in a household similar to yours. Imagine that their household is not different from other similar households, except that they like to, and do, spend more than average on [jewelry and watches]. Would you notice this about them, and if so, for how long would you have to have known them, to notice it? Would you notice it almost immediately upon meeting them for the first time, a short while after, a while after, only a long while after, or never?"*

Replies to this question were coded as 1 (almost immediately) to 5 (never). The question was repeated 32 times for each respondent, with [jewelry and watches] in the example above replaced by each of 32 expenditure category titles, randomly ordered (Heffetz, 2011). I use the same procedure as Heffetz (2011) to create the ranking of goods, i.e. I assign five values from 0 to 1 to the five response options (0, 0.25, 0.5, 0.75 and 1). I calculate the mean value for each category over all respondents.[11] The resulting visibility index ranges from 0 (least visible) to 1 (most visible).

[10] The non-response rate was 14 percent.

[11] The rankings are robust to using non-linear transformations of the data.

Moreover, I employ the consumption visibility survey developed by Charles et al. (2009). It measures the familiarity an individual needs to assess someone's above average consumption:

> *"Consider a person who lives in a household and community roughly similar to yours. How closely would you have to interact with this person in order to observe that they consistently spend more than average on each of the following categories?"*

Their answers ranged from 1 (indicating that higher than average spending could be observed if the respondent did not interact socially with the person at all) to 5 (indicating that spending would never be observed). Following Charles et al. (2009), I calculate the proportion of respondents reporting that they are able to observe above-average spending on an item even if relatively unfamiliar with the consumer.[12] The correlation of the two rankings (indices) in my dataset is 0.86 (0.84).

The results from my dataset (shown in Table 4.3) are similar to the results from the literature.[13] In particular, the most visible goods are mostly durable and non-durable goods for both of my rankings: clothes, appliances, jewelry, cars, tobacco and 'food out'. The only visible service good is education. The least visible goods tend to be services like insurance policies, legal and accounting fees, telephone charges, utilities, bills, vehicle insurance and alcohol. Yet, there are some important differences in comparison to Charles et al. (2009) and Heffetz (2011): education is perceived to be much more visible in Indonesia. Alcohol is deemed to be less visible by individuals in

[12]The results from the ranking are robust to using a linear transformation as in Heffetz (2011).

[13]Heffetz (2011) used data from a nationally representative sample from the US, whereas Charles et al. (2009) make use of a sample of business students from the University of Chicago.

Indonesia as compared to the US, which most likely results from the influences of Islam. Spending on furniture and technological devices for recreation is also considered less visible in Indonesia. These results and rankings are similar in both of my visibility rankings.

Table 4.3: Consumption Visibility rankings based on my Indonesian sample

Consumption Categories	Charles et al. (2009) Index	Charles et al. (2009) Rank	Heffetz (2011) Index	Heffetz (2011) Rank
Clothing (Clo)	.526	1	.656	1
Education (Edu)	.385	2	.562	2
Jewellery (Jwl)	.380	3	.535	5
Nonformal Education(Nonf Edu)	.361	4	.533	6
Food Out (FdO)	.350	5	.548	3
Tobacco (Cig)	.312	6	.533	7
Food at home (FdH)	.308	7	.523	8
Automobiles (Car)	.298	8	.548	3
Tvs etc. (Ot1)	.283	9	.460	13
Car Maintenance (CMn)	.280	10	.433	17
Barbershop (Brb)	.279	11	.488	9
Wedding (Wed)	.273	12	.470	12
Charity (Cha)	.265	13	.473	11
Housing (Hom)	.256	14	.342	27
Medical expenditure (Med)	.256	14	.406	20
Furniture (Fur)	.247	16	.457	15
Mobile (Cel)	.247	16	.432	18
Pilgrimage (Pil)	.245	18	.368	26
Books (Bks)	.236	19	.477	10
Alcohol Out (AlO)	.221	20	.379	24
Technological Devices (Ot2)	.219	21	.387	23
Petrol, Gas (Gas)	.219	21	.453	16
Utilities (Utl)	.203	23	.393	21
Public Transport (Bus)	.201	24	.459	14
Plane Tickets (Air)	.176	25	.391	22
Home Insurance (HIn)	.166	26	.227	31
Laundry (Lry)	.146	27	.407	19
Alcohol at Home (AlH)	.141	28	.315	28
Car Insurance (CIn)	.140	29	.276	29
Legal Fees, etc (Fee)	.140	29	.227	30
Phone Expenses (Tel)	.090	31	.370	25
Life Insurance (LIn)	.087	32	.212	32
N	115	115	115	115

In brackets next to the good categories is the abbreviation used by Heffetz (2011). A detailed description of the good categories can be found in Appendix 4.B.

Defining visible and nonvisible expenditures The complete list of expenditure categories in my consumption survey is given as follows: expenditures on rent, electricity, education, health, house maintenance, clothing, feasts and rituals, tobacco, alcohol, durable goods and household equipment, health insurance, food at home as well as miscellaneous goods (soap,

beauty, transport, reading materials, ID/driving license fee, recreation, telephone card, stamps, etc.).

My visibility indices indicated that clothing, jewelry, vehicles, education, tobacco and food out are visible goods. Only visible luxuries are hypothesised to be a signal for status. My dataset does not contain any information concerning expenditure on jewelry, vehicles and 'food out'. Therefore, I define the dependent variable as the log of expenditure on clothing and tobacco. I do not include educational expenditures into that main category as there are some concerns that the conditionalities of the cash transfer program induced mechanical increases in educational expenditures. Even though feasts do not appear as visible goods in my consumption visibility survey, anecdotal evidence from Indonesia suggests that they may be perceived as visible. In section 4.6, I show the robustness of the main results to also including feasts and educational expenditures as visible goods.

The nonvisible goods considered are given by expenditure on miscellaneous goods[14], food at home and alcohol expenses.[15] Even though food at home appears in the upper third of my consumption visibility index, I included it in the nonvisible consumption category, as it is consistently ranked as a nonvisible good in other consumption visibility surveys and papers on conspicuous consumption (Bertrand and Morse, 2016; Charles et al., 2009; Heffetz, 2011). Moreover, food expenditures are a necessity and therefore are not considered a status good. Finally, I have a third category of goods ("other") that are categorised as neither visible nor nonvisible. These goods comprise expenditures on durables, rent, electricity, education, health, house maintenance as well as feasts. Both education and health expenditure are in

[14]Soap, transport, reading materials, ID/driving license fee, recreation, telephone card, stamps.

[15]I have made sure that my results are robust to using non-food consumption shares.

this "other" category as the conditionalities of the program might mechanically induce increases in the consumption goods. Expenditure on utilities are in this category as there may be supply-driven consumption externalities. Expenditure on feasts are in this category because of direct effects of an individual's expenditure on feasts on other individuals' utility. Fourth, expenditure on housing are in this category as it is partially reliant on an estimate of rent. Lastly, durable goods are in this category as there may be important investment motives at play when purchasing durable goods.

4.4 Direct and indirect treatment effects

Specification My main specification of interest analyses consumption behaviour C_{ist}, where subscript i denotes the household, s denotes the subdistrict and t denotes the time period. I use of a difference-in-differences estimator, controlling for household fixed effects, μ_i, as well as month fixed effects, M_j. A_s denotes whether the individual lives in a treatment or control subdistrict. $POST_t$ takes value one if the observation is post-treatment and zero otherwise. I estimate the following equation using OLS:

$$C_{ist} = \theta_1 A_s \times POST_t + \theta_2 POST_t + \mu_i + M_j + \varepsilon_{ist}. \tag{4.13}$$

The main object of interest, the treatment effect, is given by θ_1. One might be concerned that unobservable sub-district level shocks result in correlations of consumption within villages or sub-districts. In addition, one could hypothesize that such shocks are also correlated across time periods. To account for this spatial correlation and the serial correlation of shocks in subdistricts over time, I cluster standard errors at the sub-district level (Angrist

and Pischke, 2008). To mitigate any concerns that the treatment effects are biased due to differences in unobservables across treatment and control subdistricts, I also report results in which I instrument the actual assignment of the treatment with the original assignment of the treatment, i.e. I identify the local average treatment effect (LATE) (Imbens and Angrist, 1994). With a randomly assigned treatment, LATE is the effect of the treatment on those who comply with the offer but are not treated otherwise.

I use a sample of 2,248 households for the analysis of the direct treatment effect and a sample of 2,542 households for the analysis of peer effects. It is reassuring that I find barely any baseline imbalances both for treated and untreated households. I find that the treated households previously received higher levels of government aid and that the untreated households display slightly higher levels of senior high school education (see Table 4.2). I also report balancing tests for a number of assets and household composition. As can be seen there are a few imbalances (see Table A.2). I have made sure that controlling for all of these covariates does not affect the results.

In the main specification I use all households from untreated villages as counterfactuals both for estimating the direct treatment effect and for estimating spillovers. I have made sure that my results are robust to conducting a propensity score exercise, that identifies households with common support. In an earlier version (Roth, 2014), I conduct a difference-in-differences strategy after dropping households without common support. The results are very similar to the ones reported in this paper. Further discussion and a detailed description of this methodology based on propensity score (and the exact list of varibles included) can be found in Roth (2014).

Identification assumptions The identifying assumptions needed to assess the causal effect of the program are given as follows:

(i) Unconfoundedness requires that conditional on observables, X_{ist}, treatment status, pkh_{ist}, is independent of potential outcomes:

$$(Y_{is0}, Y_{is1}) \perp pkh_{ist}|X_{ist}. \tag{4.14}$$

(ii) The common support assumption necessitates that for all values of the covariates in the vector of independent variables, X_{ist}, I can find households belonging to both the treatment and the control group: The baseline balance resulting from the randomisation ensures the validity of this assumption.

(iii) The stable unit treatment value assumption (SUTVA) requires that potential outcomes for household i are independent of the treatment status of any other individual k. Any potential externalities from treated to control households are thus explicitly ruled out. For the SUTVA to hold, it is crucial that the counterfactual households are taken from untreated villages. The SUTVA is not violated as I use households from untreated sub-districts as counterfactuals as long as there are no between-village spillovers. I test for the prevalence of such spillovers by using a measure of distance to the closest treated sub-district. I find no evidence for any spillovers (results available upon request).

(iv) The common trend assumption imposes a common trend for the control and the treatment group, i.e. there are no time-varying unobservables that differentially affect the treated and untreated sub-districts.

Results: Direct Effects Table 4.4 displays how people in receipt of the conditional cash transfer program adjust their consumption expenditures compared to households from control subdistricts. Total expenditure of the treated increase by roughly 9.5 percent which is slightly lower than the amount that the average household receives in a month from the conditional cash transfer program.[16] Furthermore, I find significant increases in log visible expenditure by circa 19 percent.[17] There is a statistically significant increase of nonvisible expenditure by 5.7 percent and a decrease in the expenditure share of nonvisible goods by 2 percentage points. The estimates from the local average treatment effects are very similar (Panel B).

Table 4.4: Main Results: Direct Treatment Effects

	Log total expenditure	Log Visible expenditure	Log Nonvisible expenditure	Share Visible expenditure	Share Nonvisible expenditure
Panel A: ATE					
Treatment	0.095**	0.192**	0.057*	0.001	-0.020
	(0.037)	(0.083)	(0.033)	(0.007)	(0.015)
N	4496	4197	4496	4496	4496
Panel B: LATE					
Treatment	0.087**	0.175**	0.064*	-0.001	-0.013
	(0.039)	(0.083)	(0.035)	(0.007)	(0.016)
N	4496	4197	4496	4496	4496
Number of Clusters	85	85	85	85	85
Household FE	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y

This table uses treated households from treatment villages and all households from control villages. Standard errors clustered at the sub-district level in parentheses. All specifications control for household fixed effects, month fixed effects, and a $POST_t$ indicator taking value one in the post-treatment period. Panel A shows the average treatment effects (ATE) using the actual treatment assignment. Panel B shows the local average treatment effect (LATE), instrumenting people's actual treatment assignment with the original treatment assignment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

[16]This finding is in line with a life-cycle model of consumption with credit-constrained households. It is not compatible with a classical life-cycle model without credit-constraints, in which individuals would be expected to smooth their consumption inter-temporally.

[17]Less than 11 percent of households reported 0 spendings on visible goods at either the baseline or follow-up. These observations are missing in the analyses based on logs. I demonstrate that my results are robust in levels.

Results: Indirect Effects While the previous results highlight that the cash transfer increased the visible expenditures of treated households, Panel A of Table 4.5 analyzes the consumption behaviour of untreated households in treated subdistricts compared to households from control subdistricts. This allows us to study spillovers of the cash transfer program. Untreated households in treated subdistricts significantly increase their visible expenditures by 25 percent, and significantly shift their consumption shares: visible expenditure shares increase by two percentage points. I find a statistically insignificant decrease in expenditure on nonvisible goods by three percent which corresponds to a statistically significant decrease in nonvisible expenditure shares by three percentage points. Yet, I find no evidence of significant increases in overall expenditure of the untreated households in treated sub-districts. Log total expenditure increase in a statistically insignificant manner by 3.5 percent. The results based on the local average treatment effect are similar to the results from above (see Panel B). The patterns in the data are consistent with the model of relative concerns over visible goods as presented in section 4.2.2 and models of social learning (Bikhchandani et al., 1992), but inconsistent with the signalling model of conspicuous consumption outlined in Section 4.2.1.

4.5 Mechanisms

4.5.1 Social mechanisms

In this section, I use several proxies for how the degree of anonymous interactions in a given village affects consumption responses to the cash transfer program. I make use of baseline data on membership in social organisations

Table 4.5: Spillover effects on ineligible households

	Total log expenditure	Log Visible expenditure	Log Nonvisible expenditure	Share Visible expenditure	Share Nonvisible expenditure
Panel A: ATE					
Treatment	0.035	0.254***	-0.017	0.021**	-0.032**
	(0.041)	(0.085)	(0.033)	(0.008)	(0.014)
N	5084	4705	5084	5084	5084
Panel B: LATE					
Treatment	0.007	0.188**	-0.023	0.015*	-0.021
	(0.045)	(0.082)	(0.037)	(0.008)	(0.016)
N	5084	4705	5084	5084	5084
Number of Clusters	85	85	85	85	85
Household FE	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y

This table uses untreated households from treatment villages and all households from control villages. Standard errors clustered at the sub-district level in parentheses. All specifications control for household fixed effects, month fixed effects, and a $POST_t$ indicator taking value one in the post-treatment period. Panel A shows the average treatment effects (ATE) using the actual treatment assignment. Panel B shows the local average treatment effect (LATE), instrumenting people's actual treatment assignment with the original treatment assignment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

(religious study groups, neighbourhood associations, and women's groups, but also on many self-help organisations), village population size, social punishment norms[18], and ethnic heterogeneity. I focus on results using a sample-split method and employ the same specification as for the analysis above. For simplicity's sake, this section focuses on changes in consumption shares of visible and nonvisible goods.

Effect on the treated First, I examine heterogeneity of consumption responses of the treated. The increases in visible expenditure of the treated are more pronounced for households with lower baseline social activities, in villages without social punishment norms and with lower social activities, higher ethnic fractionalisation, and higher population size (Table 4.6).

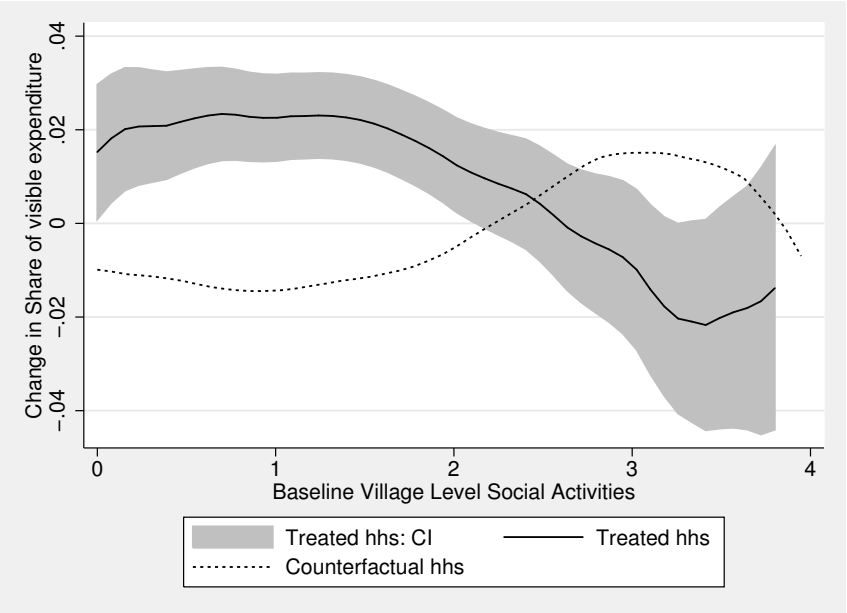
[18]The village head is asked whether households are punished when they do not contribute to community projects.

The increases in visible expenditure are only statistically significant in villages without social punishment norms which proxy for social sharing constraints. These norms could operate in the following way: if only village heads and treated households know about the conditional cash transfer program, then visible consumption could give away that a household received the cash transfer. The transfer might then be more heavily taxed in villages with social punishment norms.

Spillover effects In what follows, I examine heterogeneity of peer effects by several measures of social activities. The spillover effects on visible expenditure shares are lower in villages with more social activities and for households with more social activities. Figure 4.1 and Table 4.6 show heterogeneous treatment effects of visible consumption shares by baseline village social club memberships. The increases in visible expenditure for the ineligibles are completely driven by individuals from villages with lower social activities, consistent with the hypothesis by Veblen (1899) that in societies with more anonymous interactions, conspicuous consumption is more prevalent.

I further shed light on the role of social capital. In line with the previous results on social capital, I find that the peer effects in visible consumption are significantly stronger for households in villages with a larger population (see Panel A of Table 4.7). Furthermore, I use an alternative proxy for social capital. Specifically, I use data on ethnic fractionalisation, a commonly used correlate of social capital. There is substantial evidence that levels of social capital are negatively correlated with ethnic diversity (Alesina and La Ferrara, 2000). In districts with higher levels of ethnic fractionalisation, increases in visible consumption shares are more pronounced both for

Figure 4.1: Treatment Heterogeneity by Baseline Village Level Social Activities on Ineligibles: Visible Consumption shares



This figure displays a local polynomial regression of the change in visible expenditure share between the baseline and post-treatment period and the average number of social organisations at the village. The figure shows these local polynomial regresssion separately for ineligilble households in treatment subdistricts and for counterfactual households from control subdistricts.

treated and, in particular, untreated households (see Panel D of Table 4.6).

The observed treatment heterogeneity by social capital seems at odds with an alternative explanation in which people share their income or their consumption with their neighbours. Specifically, if sharing is more prevalent in places with more social activities, then we would expect the spillover effects to be larger. Yet, this is the exact opposite of what I observe in the data. In the following subsection, I provide further evidence on the irrelevance of sharing for explaining the patterns in the data.[19]

[19]My results on treatment heterogeneity are robust to conducting a triple-differences analysis, i.e. interacting baseline values of the variables of interest with a treatment-post-indicator, while also interacting the variable of interest with a post-indicator. These results are reported in Table A.3. Moreover, in an earlier version (Roth, 2014), I show that my main results on treatment heterogeneity by social capital and social punishment norms are robust to the inclusion of a variety of interaction terms at once.

Table 4.6: Heterogeneity: Sample Split - 1

	Treatment Effect on Eligibles				Treatment Effect on Ineligibles			
	Share Visible		Share Nonvisible		Share Visible		Share Nonvisible	
Panel A								
High Social Activities (HH level)	Y	N	Y	N	Y	N	Y	N
Treatment	-0.005	0.007	-0.011	-0.031	0.010	0.032***	-0.017	-0.054***
	(0.010)	(0.009)	(0.020)	(0.019)	(0.011)	(0.009)	(0.022)	(0.017)
P-value (equality)	0.089*		0.248		0.043**		0.079*	
N	2070	2426	2070	2426	2362	2722	2362	2722
Panel B								
High Social Activities (village level)	Y	N	Y	N	Y	N	Y	N
Treatment	-0.004	0.010	-0.013	-0.028	0.011	0.038***	-0.024	-0.053***
	(0.010)	(0.010)	(0.022)	(0.018)	(0.012)	(0.010)	(0.022)	(0.018)
P-value (equality)	0.127		0.292		0.035**		0.212	
N	2312	2332	2312	2332	2552	2744	2552	2744
Panel C								
Punishment	Y	N	Y	N	Y	N	Y	N
Treatment	-0.015	0.015*	-0.024	-0.022	0.006	0.033***	-0.023	-0.042**
	(0.012)	(0.009)	(0.022)	(0.018)	(0.015)	(0.009)	(0.024)	(0.018)
P-value (equality)	0.020**		0.941		0.092*		0.465	
N	1942	2554	1942	2554	1980	3104	1980	3104
Panel D								
High Ethnic Fractionalisation	Y	N	Y	N	Y	N	Y	N
Treatment	0.013	-0.010	-0.044*	-0.001	0.038***	0.007	-0.059**	-0.016
	(0.011)	(0.010)	(0.023)	(0.020)	(0.011)	(0.013)	(0.024)	(0.022)
P-value (equality)	0.125		0.163		0.062*		0.183	
N	2256	2240	2256	2240	2518	2566	2518	2566
Household Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y

Standard errors are clustered at the sub-district level. For this sample split method I bifurcate the data. In Panel A, those with high social activities are those with at least two social club memberships. In Panel B, those with high social capital are those living in villages with at least 1.5 social club memberships on average. In Panel C, I bifurcate the data into villages with and without punishment as reported by village heads. In Panel D, I bifurcate the data by median ethnic fractionalisation. All of these specifications include households fixed effects. The equality of coefficient estimates is tested based on seemingly unrelated regressions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.5.2 Alternative mechanisms

In what follows, I rule out explanations based on sharing norms and human capital motives as well as local price effects (proxied by distance to the district capital) as alternative mechanisms. The results are displayed in Table 4.7, panels B to D. First, I use several variables from the culture module of the Indonesian socioeconomic survey to examine heterogeneous treatment effects by sharing norms.[20] If my results were driven by sharing, the spillovers on ineligibles would be larger in areas with stronger sharing norms. I do not find any evidence for heterogeneous reponses by sharing norms. Second, I investigate treatment heterogeneity by educational achievement of the household head. Moav and Neeman (2012) develop a theoretical model that predicts higher levels of conspicuous consumption for the less educated. I find no evidence of heterogeneous effects by the level of schooling (see Panel C of Table 4.7). Third, I find no treatment heterogeneity by distance to the district capital, proxying for market integration (see Panel D of Table 4.7). This suggests that local price increases are not driving my results.

4.6 Robustness

In this Section, I examine the robustness of results and asses various alternative mechanisms, such as income effects, non-linear Engel curves and local price effects. Finally, I demonstrate robustness of results by estimating different model specifications.

[20]This data on food sharing norms is only available with subdistrict identifiers in the 2009 wave of the SUSENAS. Moreover, this dataset is only available for about 90 percent of the subdistricts in my sample.

Table 4.7: Heterogeneity: Sample Split -2

	Treatment Effect on Eligibles				Treatment Effect on Ineligibles			
	Share Visible		Share Nonvisible		Share Visible		Share Nonvisible	
Panel A								
High Population Size	Y	N	Y	N	Y	N	Y	N
Treatment	0.007	-0.004	-0.025	-0.020	0.032***	0.012	-0.046**	-0.028
	(0.007)	(0.012)	(0.019)	(0.021)	(0.010)	(0.011)	(0.021)	(0.020)
P-value (equality)	0.413		0.841		0.082*		0.462	
N	2254	2242	2254	2242	2552	2532	2552	2532
Panel B								
High Food Sharing	Y	N	Y	N	Y	N	Y	N
Treatment	0.007	-0.003	-0.016	-0.031	0.024*	0.022**	-0.043*	-0.034
	(0.013)	(0.007)	(0.022)	(0.022)	(0.013)	(0.011)	(0.025)	(0.020)
P-value (equality)	0.512		0.628		0.903		0.769	
N	2416	2080	2416	2080	2616	2468	2616	2468
Panel C								
No Schooling	Y	N	Y	N	Y	N	Y	N
Treatment	0.007	-0.001	-0.025	-0.021	0.020**	0.023**	-0.042**	-0.034*
	(0.011)	(0.009)	(0.024)	(0.015)	(0.010)	(0.010)	(0.020)	(0.019)
P-value (equality)	0.529		0.832		0.753		0.706	
N	1692	2804	1692	2804	1850	3234	1850	3234
Panel D								
High Distance to District Capital	Y	N	Y	N	Y	N	Y	N
Treatment	0.009	-0.006	-0.040*	-0.003	0.030***	0.014	-0.042**	-0.033
	(0.010)	(0.011)	(0.020)	(0.022)	(0.011)	(0.013)	(0.020)	(0.024)
P-value (equality)	0.319		0.188		0.352		0.756	
N	2368	2128	2368	2128	2614	2470	2614	2470
Household Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y

Standard errors are clustered at the sub-district level. In Panel A, I split the sample into individuals living in villages above and below 4030 individuals. In Panel B, I bifurcate the data by the median (2.93) agreement to a question on food sharing on a four point likert scale. In Panel C I bifurcate the data by information on whether an individual received any schooling. In Panel D, I bifurcate the data into villages with a distance to the district capital below and above 20 kilometers. All of these specifications include households fixed effects. The equality of coefficient estimates is tested based on seemingly unrelated regressions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Income effects and non-linear Engel curves Total expenditure of the untreated households did not increase in a significant manner. Yet, the point estimates indicate a statistically insignificant increase of total expenditure by 3.5 percentage points. Could this small increase in total expenditure of the untreated explain their change in expenditure shares? Visible goods are luxuries. Therefore, it would be possible that the increase in visible expenditure shares of the untreated households could be explained by the increase in total expenditure, i.e. by a "luxury-effect". In particular, the increased overall expenditure could explain the increase in visible expenditure if the estimated elasticity of visible expenditure with respect to total expenditure is sufficiently high. I use the treatment effect on log expenditure and log visible consumption to calculate the implied elasticities of visible expenditure with respect to total expenditure.[21] I find that for the treated the implied elasticity (standard error in brackets) is given by 1.65 (.55). The implied elasticity of visible expenditure with respect to total expenditure for the untreated households is given by 9.17 (9.85). To rationalise these results, the preferences of treated and untreated households would have to differ substantially.

In what follows, I rule out that non-linear Engel curves can account for the patterns in my data. In particular, Engel curves of visible expenditure shares that decrease at a certain total expenditure level could potentially explain the patterns in the data. Figure 4.2 depicts how a non-linear hump-shaped Engel curve could explain the patterns in the data.[22]

I examine the shape of the Engel curve which is estimated non-parametrically

[21] To calculate the elasticities I restrict the sample to those households with positive visible expenditure.

[22] The baseline mean log expenditure of treated and untreated household are denoted by BT and BU respectively. Similarly, the follow-up mean log expenditure of the untreated and treated are denoted by FU and FT respectively.

by local polynomial kernel regressions and is displayed in Figure 4.3. The Figure highlights that the Engel curve is approximately linear.

Figure 4.2: Non-linear Engel curve of visible expenditure shares

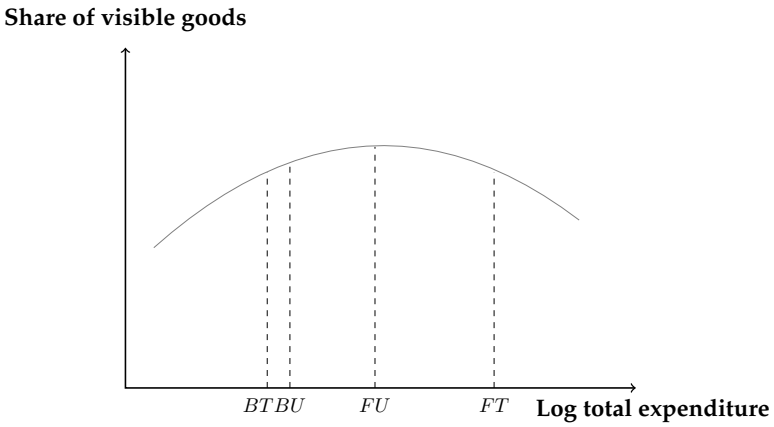
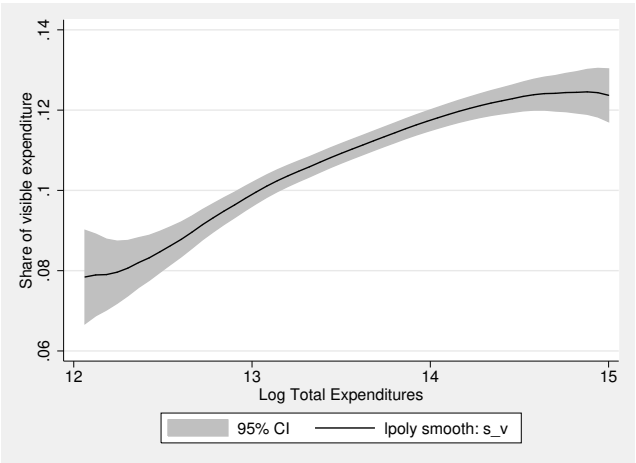


Figure 4.3: Engel Curve of Visible expenditure Shares



This figure shows local polynomial regressions of visible expenditure shares on the log of total expenditures. I removed the top and bottom one percent of the expenditure distribution to mitigate the influence of outliers

Local price and supply side effects Can changes in local prices or local supply side conditions explain the patterns in the data? The cash transfer program might result in local price increases. Thus, what seems like a sub-

stantial increase in expenditure on a certain good would just reflect differential price effects between treatment and control sub-districts. In what follows, I provide evidence on the irrelevance of local price effects: First, there is evidence that conditional cash transfer programs do not result in significant general equilibrium effects in the form of prices and wages (Angelucci and De Giorgi, 2009; Fiszbein et al., 2009). Even though such large differential price effects for visible goods seem rather unlikely, I demonstrate that the implied supply elasticities needed for the price effects to explain a large fraction of the effects on ineligible households are implausible.

I engage in the following thought experiment: how low would the supply elasticity have to be to explain the indirect treatment effects through prices? I calculate the implied supply elasticity for visible expenditure based on (a) the observed change in demand for visible expenditure (given by the treatment effect on eligibles) and (b) the observed visible expenditure changes from the untreated households (given by the treatment effect on ineligibles). If untreated households have the same expenditure bundles as at baseline, but the results are completely driven by price effects, then the implied supply elasticity (with the standard error[23] in brackets) for visible expenditure is:

$$\text{supply elasticity} = \frac{\text{change in demand}}{\triangle \text{price}} = 0.076 \ (0.025). \tag{4.15}$$

As the visible consumption of about 10 percent of the population increases by almost 20 percent, overall demand for visible goods in the village increases by at most 2 per cent, while visible expenditure of untreated households increase by roughly 25 percent. The implied supply elasticity of 0.076

[23]To recover an estimate of the supply elasticity (and its standard error), I use the estimates of the visible expenditure increases for the treated and untreated.

seems implausibly low.

Further, to account for local price effects, I deflate prices by the district-specific poverty lines constructed by the Indonesian statistical agency (BPS). District poverty lines evaluate the cost of a food basket and a non-food basket that is deemed to correspond to meeting basic food and non-food needs. Results are robust to deflating expenditure by district-level poverty lines (Panel H in Table 4.8). Once prices are deflated I find no more evidence at all of an increase in total expenditure of the untreated. Yet, the effect on visible expenditure shares and log-visible expenditure remains.

To rule out that supply-side changes drive the results, I use data from the census of villages (PODES) from 2005 prior to the treatment and after the treatment (in 2008) to show that the supply environment was not significantly affected by the receipt of the cash transfer program. In particular, I use data on the number of markets, shopping areas, agricultural kiosks, small industry firms, farming businesses and mini markets in the village as well as information on the distance to the closest shopping area and market. I find neither baseline differences in market environments between treated and untreated sub-districts nor any significant treatment effects (Table A.7).

Finally, local price effects would be expected to be highest in the smallest villages as those are likely least integrated with other markets. Therefore, if local price effects are behind the estimated treatment effects, they would be the largest in small villages. Yet, I find that the treatment effects are smaller in relatively smaller villages. All in all, this suggests that local price effects are very unlikely driving the observed spillover effect.

Accounting exercise Do the budget identities hold? In this subsection, I split the budget into visible goods, nonvisible goods and other consumption. First, I show that the sum of visible, nonvisible and other consumption adds up to the total expenditure (Panel A of Table A.4). Second, I show that the budget identity also holds for untreated households in treated villages (Panel B of Table A.4).

Further robustness checks I demonstrate robustness of results to interacting baseline controls with a post-indicator (Panel D of Table 4.8), excluding outliers (Panel E), interacting proxies for supply side conditions with a post-indicator (Panel F), clustering standard errors at the vilage level (Panel G) and deflating consumption by district level poverty lines (Panel H). Results are also robust to clustering standard errors at the district level (Panel B in Table A.5)[24], excluding month fixed effects (Panel C in Table A.5), and analyzing consumption in levels rather than logs (Panel D in Table A.5).

In addition, I conduct an analysis for all consumption goods separately: my main results for the untreated are mostly driven by expenditures on clothing. As clothing is the most visible good in my consumption visibility survey this result is supportive of the visibility mechanism (see Table A.9). However, the treatment effects on consumption expenditures in levels are all quite noisily measured. Finally, I show that my results are robust to including educational expenditure as well as expenditure on feasts in my visible consumption category. Specifically, I use four different definitions of visible goods (see Table A.6), which also include spending on feasts and education.

[24] As there are only 28 districts, I have made sure that my results are robust to conducting a wild-cluster bootstrap correction (Cameron and Miller, 2015).

Table 4.8: Robustness

	Treatment Effect on Eligibles					Treatment Effect on Ineligibles				
	Total expenditure	Visible expenditure	Nonvisible expenditure	expenditure Share Visible	Nonvisible	Total expenditure	Visible expenditure	Nonvisible expenditure	expenditure Share Visible	Nonvisible
Panel A: Main Specification										
Treatment	0.095**	0.192**	0.057*	0.001	-0.020	0.035	0.254***	-0.017	0.021**	-0.032**
	(0.037)	(0.083)	(0.033)	(0.007)	(0.015)	(0.041)	(0.085)	(0.033)	(0.008)	(0.014)
Panel B: Subdistrict FE										
Treatment	0.098***	0.188**	0.058*	0.001	-0.021	0.037	0.254***	-0.016	0.021**	-0.032**
	(0.037)	(0.083)	(0.033)	(0.007)	(0.015)	(0.041)	(0.085)	(0.033)	(0.008)	(0.014)
Panel C: District Trends										
Treatment	0.108***	0.225**	0.058*	0.003	-0.023*	-0.014	0.202***	-0.036	0.019***	-0.016
	(0.035)	(0.086)	(0.030)	(0.007)	(0.012)	(0.038)	(0.069)	(0.029)	(0.005)	(0.010)
Panel D: Controls × Post										
Treatment	0.081**	0.161*	0.053	-0.000	-0.015	0.031	0.242***	-0.018	0.020***	-0.031**
	(0.036)	(0.085)	(0.032)	(0.007)	(0.015)	(0.040)	(0.083)	(0.032)	(0.008)	(0.014)
Panel E: No outliers										
Treatment	0.073**	0.165**	0.038	0.001	-0.017	0.024	0.218***	-0.025	0.017**	-0.032***
	(0.032)	(0.078)	(0.032)	(0.007)	(0.012)	(0.033)	(0.076)	(0.031)	(0.007)	(0.011)
Panel F: Supply-Side × Post										
Treatment	0.073**	0.139*	0.042	0.002	-0.015	0.021	0.221***	-0.031	0.019***	-0.031**
	(0.034)	(0.076)	(0.034)	(0.006)	(0.014)	(0.035)	(0.072)	(0.031)	(0.007)	(0.013)
Panel G: Cluster at the Village-Level										
Treatment	0.095***	0.192***	0.057*	0.001	-0.020*	0.035	0.254***	-0.017	0.021***	-0.032***
	(0.031)	(0.068)	(0.030)	(0.006)	(0.011)	(0.031)	(0.066)	(0.027)	(0.006)	(0.011)
Panel H: Deflated by Poverty lines										
Treatment	0.036	0.121	0.005	0.002	-0.016	-0.005	0.201**	-0.058	0.018**	-0.031***
	(0.041)	(0.079)	(0.038)	(0.006)	(0.012)	(0.042)	(0.079)	(0.036)	(0.007)	(0.011)
N	4496	4197	4496	4496	4496	5084	4705	5084	5084	5084

All specifications include a time trend. Except for Panel B which contains subdistrict fixed effects all specifications include household fixed effects. Standard errors are clustered at the sub-district level in all Panels except for G. In Panel C I include district trends. In Panel D I include the following baseline covariates interacted with a post-dummy: number of social clubs, age, households size, a dummy for no schooling and junior high school, whether the households works in agriculture, received aid beforehand and several variables that measure distance to markets. In Panel E I exclude the bottom and top 1 percent of the expenditure dstribution. In Panel F I include supply-side controls interacted with a post-indicator. In Panel G I cluster standard errors at the village level. In Panel H I deflate the consumption data by district specific poverty lines.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.7 Conclusion

This paper provides evidence that people's share of visible consumption increase in peer group income. Using data on poor households from Indonesia I show that the consumption share of visible expenditures increases by 20 percent in response to an increase in peer group income. These patterns are consistent with a simple model of relative concerns over visible goods, but inconsistent with a standard signalling model of conspicuous consumption.

Moreover, I shed light on the social mechanisms underlying the tendency to consume visible goods. In villages with more social interactions there is almost no increase in visible expenditure of the untreated. On the other hand, in villages with low social activities there is a strong increase in visible expenditure for the untreated. Interestingly, this evidence is consistent with the idea by Veblen (1899) that conspicuous consumption is higher in environments with more "anonymous interactions" and lower "social capital".

Peer effects in consumption affect optimal policy design. While Angelucci and De Giorgi (2009) document potentially welfare-enhancing spillovers of a conditional cash transfer program through sharing, my paper emphasizes potentially welfare-reducing spillovers as a result of status concerns. The design of a conditional cash transfer program, for example, could depend on the prevalence and size of peer effects for several reasons. Therefore, the conditionalities imposed by conditional cash transfer programs could be guided by empirical findings on spillovers in consumption behavior.

Appendices of Chapter 4

4.A Additional Tables

Table A.1: Baseline characteristics of households in treated and untreated sub-districts in the original treatment assignment

	Treated Sub-districts	Control Sub-districts	P-value: Equality
Average Consumption			
Total expenditure	97.491	97.546	.991
Log Visible expenditure	11.16	11.223	.471
Visible expenditure	10.36	10.81	.568
Nonvisible expenditure	58.357	60.276	.492
Log Nonvisible expenditure	13.16	13.19	.568
Share Visible expenditure	0.108	0.112	.467
Share Nonvisible expenditure	0.672	0.665	.563
Average Demographics (Household Head)			
Main Profession: Agriculture	0.624	0.561	.163
Household Size	5.082	5.131	.734
Age of Household Head	44.70	44.48	.742
No Schooling	0.358	0.301	.119
Junior High School	0.092	0.111	.193
Senior High School	0.049	0.052	.716
Previous Government Aid	0.944	0.915	.042**
Connection to Village Head	0.529	0.506	.668
No Sanction	0.645	0.689	.506
Average Village Characteristics			
Percentage Islam	93.56	93.61	.991
Population Size	4858	4551	.556
Percentage of Farmers	70.17	66.11	.161
Percentage without Electricity	0.023	0.020	.683
Ethnic Fractionalisation: Herfindhal Index	0.146	0.130	.668
Village Head Sanction	0.38	0.42	.497
Observations	1524	1796	

The expenditure in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar. Standard errors underlying the tests for the equality of means are clustered at the subdistrict level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Baseline characteristics of households of treated and untreated households in treated sub-districts and households from untreated villages: Actual assignment

Panel	A	B	C	P-value	
	Direct: Treatment	Indirect: Treatment	Control	A-B	A-C
Radio	0.415	0.452	0.456	0.211	0.879
Bicycle	0.592	0.678	0.618	0.710	0.342
Motorcycle	0.110	0.230	0.187	0.015**	0.189
Car	0.003	0.003	0.003	0.724	0.785
Handphone	0.074	0.116	0.097	0.196	0.250
Chicken	0.366	0.423	0.415	0.131	0.833
Pig	0.009	0.011	0.005	0.607	0.439
Goat	0.094	0.085	0.127	0.082*	0.172
Cow	0.035	0.062	0.053	0.288	0.670
Age between 0 and 2	0.408	0.385	0.402	0.805	0.362
Age between 3 and 6	0.583	0.460	0.508	0.036**	0.102
Age between 7 and 15	1.277	1.089	1.262	0.849	0.018**
Age between 16 and 24	0.572	0.693	0.635	0.251	0.306
Age between 25 and 39	1.194	1.174	1.185	0.813	0.779
Age between 40 and 59	0.875	0.928	0.914	0.336	0.744
Age older than 60	0.212	0.291	0.257	0.116	0.224
Observations	773	1067	1475	2248	2542

Standard errors underlying the tests for the equality of means are clustered at the subdistrict level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Treatment Heterogeneity

	Direct Treatment Effect on the treated					Peer Effect (indirect treatment effect)				
	Total expenditure	Visible expenditure	Nonvisible expenditure	expenditure Share Visible	Nonvisible	Total expenditure	Visible expenditure	Nonvisible expenditure	expenditure Share Visible	Nonvisible
Panel A:										
Treatment	0.064	0.252**	0.030	0.015*	-0.022	0.020	0.312***	-0.035	0.033***	-0.042**
	(0.054)	(0.105)	(0.051)	(0.009)	(0.018)	(0.052)	(0.103)	(0.045)	(0.009)	(0.018)
Treatment ×	0.094	-0.125	0.070	-0.030**	-0.002	0.061	-0.111	0.050	-0.027*	0.019
Sanction	(0.071)	(0.167)	(0.065)	(0.013)	(0.026)	(0.064)	(0.164)	(0.059)	(0.016)	(0.026)
Panel B:										
Treatment	0.094*	0.249*	0.028	0.012	-0.041*	0.038	0.431***	-0.060	0.041***	-0.064***
	(0.055)	(0.131)	(0.056)	(0.011)	(0.021)	(0.054)	(0.110)	(0.051)	(0.010)	(0.018)
Treatment ×	0.005	-0.034	0.020	-0.007	0.013	0.007	-0.103***	0.028	-0.013***	0.018**
Household Social Activity	(0.023)	(0.055)	(0.023)	(0.005)	(0.009)	(0.022)	(0.038)	(0.020)	(0.004)	(0.008)
Panel C:										
Treatment	0.080	0.231	0.006	0.016	-0.047*	0.040	0.536***	-0.077	0.050***	-0.081***
	(0.078)	(0.163)	(0.084)	(0.014)	(0.026)	(0.078)	(0.140)	(0.080)	(0.014)	(0.028)
Treatment ×	0.015	-0.023	0.034	-0.009	0.015	0.006	-0.174**	0.041	-0.019**	0.029*
Village Social Activities	(0.038)	(0.085)	(0.042)	(0.008)	(0.015)	(0.041)	(0.076)	(0.041)	(0.009)	(0.017)
Panel D:										
Treatment	0.368	0.114	0.331	-0.125	0.005	0.053	-0.487	-0.084	-0.143*	-0.022
	(0.365)	(1.038)	(0.391)	(0.082)	(0.150)	(0.410)	(0.970)	(0.393)	(0.081)	(0.164)
Treatment	-0.032	0.010	-0.033	0.015	-0.003	-0.000	0.092	0.009	0.020**	-0.002
Population Size	(0.044)	(0.121)	(0.047)	(0.010)	(0.018)	(0.049)	(0.117)	(0.047)	(0.010)	(0.020)
Panel E:										
Treatment	0.183	-0.330	0.138	-0.038	-0.052	0.029	-0.014	0.032	0.009	-0.034
	(0.289)	(0.429)	(0.285)	(0.038)	(0.088)	(0.293)	(0.445)	(0.259)	(0.047)	(0.099)
Treatment ×	-0.027	0.181	-0.028	0.014	0.009	0.008	0.099	-0.017	0.005	-0.002
Food Sharing Norms	(0.096)	(0.149)	(0.097)	(0.013)	(0.031)	(0.102)	(0.157)	(0.091)	(0.016)	(0.035)
Panel F:										
Treatment	0.101	0.158	0.100*	-0.010	-0.001	0.014	0.085	-0.014	0.007	-0.016
	(0.063)	(0.136)	(0.057)	(0.010)	(0.020)	(0.062)	(0.125)	(0.048)	(0.013)	(0.021)
Treatment ×	0.006	0.080	-0.078	0.023	-0.043	0.072	0.366**	-0.003	0.031*	-0.042
Ethnic Heterogeneity	(0.083)	(0.185)	(0.079)	(0.015)	(0.031)	(0.087)	(0.173)	(0.072)	(0.017)	(0.032)
Panel G:										
Treatment	0.028	0.083	0.001	-0.014	-0.005	0.042	0.163	-0.017	0.008	-0.024
	(0.075)	(0.169)	(0.080)	(0.014)	(0.027)	(0.084)	(0.173)	(0.074)	(0.017)	(0.030)
Treatment ×	0.003	0.005	0.003	0.001	-0.001	0.000	0.005	0.000	0.001	-0.001
Distance to District Capital	(0.003)	(0.007)	(0.003)	(0.001)	(0.001)	(0.003)	(0.007)	(0.003)	(0.001)	(0.001)
Household Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N	4496	4036	4496	4496	4496	5084	4384	5084	5084	5084

Standard errors are clustered at the sub-district level. Baseline covariates are interacted with the treatment-post indicator. All specifications include a time trend and control for differential trend by the interaction term, i.e. all interaction terms are multiplied with a Post indicator. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: Does the budget constraint bind?

	Expenditures in levels			
	Total Expenditures	**Visible Expenditures**	**Nonvisible Expenditures**	**Other consumption**
	Panel A: Treatment Effect on the treated			
Treatment	95.71*	18.26	33.58	43.85
	(49.04)	(11.00)	(25.27)	(35.78)
N	4495	4495	4495	4495
	Panel B: Treatment Effect on Ineligibles (Peer Effect)			
Treatment	70.00	34.17**	-17.13	-10.03
	(66.08)	(13.24)	(23.40)	(53.70)
N	5082	5082	5082	5082
Household FE	Y	Y	Y	Y
Month FE	Y	Y	Y	Y

Standard errors clustered at the subdistrict level in parentheses. The expenditure in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar..
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: Further Robustness of Treatment Effects

	Direct Treatment Effect on the treated					Indirect Treatment Effect on the untreated				
	Total expenditure	Visible expenditure	Nonvisible expenditure	expenditure Share Visible	Nonvisible	Total expenditure	Visible expenditure	Nonvisible expenditure	expenditure Share Visible	Nonvisible
Panel A: Main specification										
Treatment	0.095**	0.192**	0.057*	0.001	-0.020	0.035	0.254***	-0.017	0.021**	-0.032**
	(0.037)	(0.083)	(0.033)	(0.007)	(0.015)	(0.041)	(0.085)	(0.033)	(0.008)	(0.014)
Panel B: Cluster at the District-Level										
Treatment	0.095***	0.192***	0.057*	0.001	-0.020	0.095***	0.192***	-0.017	0.021***	-0.032**
	(0.034)	(0.062)	(0.031)	(0.005)	(0.014)	(0.034)	(0.062)	(0.032)	(0.007)	(0.015)
Panel C: Without Month FE										
Treatment	0.104**	0.197**	0.060	0.002	-0.023	0.050	0.271***	-0.015	0.022**	-0.037**
	(0.041)	(0.093)	(0.039)	(0.008)	(0.016)	(0.044)	(0.090)	(0.036)	(0.009)	(0.016)
Panel D: In levels										
Treatment	59.60	14.09	21.71	0.002	-0.016	12.18	31.47**	-25.45	0.018**	-0.031***
	(45.21)	(11.27)	(25.97)	(0.006)	(0.012)	(53.55)	(13.39)	(23.19)	(0.007)	(0.011)
N	4496	4050	4496	4496	4496	4496	4050	5084	5084	5084

All specifications include a time trend and household fixed effects. All specifications include month fixed effects except for the one displayed in Panel C. Panel D shows treatment effects in levels. The expenditure in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar. Standard errors are clustered at the sub-district level in all Panels except for Panel B, where they are clustered at the district-level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6: Different Definitions of Visible Goods

Panel	(A)	(B)	(C)	(D)
		Treatment Effect on Eligibles		
Treatment	0.192**	0.139	0.168*	0.173*
	(0.083)	(0.105)	(0.100)	(0.089)
N	4050	4290	4255	4112
		Treatment Effect on Ineligibles		
Treatment	0.254***	0.139	0.180*	0.200**
	(0.085)	(0.100)	(0.093)	(0.096)
N	4506	4807	4764	4580
Household Fixed Effects	Y	Y	Y	Y

Standard errors clustered at the sub-district level in parentheses. All specifications include a time trend and household fixed effects. All specifications control for month fixed effects. In column A, I display the results from the main specification. In Panel B, I use expenditure on education, tobacco, clothing and feasts for my main specification. In Panel C, I use expenditure on education, tobacco and clothing. In Panel D, I use expenditure on feasts, tobacco and clothing. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Supply-Side Characteristics at the Village Level

	Deflator	**Production Kiosk**	**Small Industry**	**Shopping Area**
Treatment	0.000	-0.240	6.471	-0.001
	(0.000)	(0.194)	(12.386)	(0.019)
Treatment × Post	0.045	-2.465	35.245	-0.370
	(0.034)	(2.933)	(117.709)	(0.305)
Number of Villages	1304	1272	1272	1272
	Distance to Shopping	**Market**	**Distance to Market**	**Number of Minimarkets**
Treatment	0.495	0.028	0.302	-0.032
	(0.700)	(0.028)	(0.555)	(0.034)
Treatment × Post	-8.110	0.249	-3.815	0.022
	(14.682)	(0.370)	(5.674)	(0.370)
Number of Villages	1272	1272	1272	1272
Time dummy	yes	yes	yes	yes

Standard errors clustered at the sub-district level in parentheses. All specifications include a time trend. For a small number of villages (32 out of 1310) I was not able to match the supply-side data. * $p < 0.10$, ** $p < 0.05$, ***

Table A.8: Baseline characteristics of contaminated vs. uncontaminated subdistrics.

	Contaminated Sub-districts	Uncontaminated Sub-districts	P-value
Average Consumption			
Log Total expenditure	13.59	13.63	0.628
Log Visible expenditure	11.18	11.20	.862
Log Nonvisible expenditure	13.14	13.19	.576
Share Visible expenditure	0.11	0.11	.813
Share Nonvisible expenditure	0.67	0.67	.872
Average Demographics (Household Head)			
Main Profession: Agriculture	0.49	0.60	.031**
Household Size	4.97	5.12	.244
Age of Household Head	44.05	44.64	.695
No Schooling	0.24	0.34	.025**
Junior High School	0.14	0.10	.099*
Senior High School	0.09	0.05	.083*
Previous Government Aid	0.91	0.93	.295
Connection to Village Head	0.53	0.51	.806
No Sanction	0.67	0.67	.955
Average Village Characteristics			
Percentage Islam	86.59	94.33	.491
Population Size	4173	4747	.256
Percentage of Farmers	60.42	68.77	.878
Percentage without Electricity	0.02	0.02	
Ethnic Fractionalizarion: Herfindhal Index	0.15	0.14	.921
Village Head Sanction	0.44	0.40	.689
Observations	316	3004	3320

Contaminated subdistricts are those whose actual treatment assignment differs from the original treatment assignment. The expenditure in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar. Standard errors underlying the tests for the equality of means are clustered at the subdistrict level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Individual Consumption Categories in levels

	Clothing	Tobacco	Food	Miscellaneous Goods
		Panel A: Treatment Effect on Eligibles		
Treatment	13.23	34.03	40.94*	-7.6
	(8546.367)	(5489.865)	(22642.209)	(6907.272)
N	4496	4496	4493	4495
		Panel B: Treatment Effect on Ineligibles		
Treatment	23.80**	11.29*	-14.03	-2.97
	(10727.179)	(5890.836)	(20519.567)	(6617.612)
N	5081	5083	5080	5083

Standard errors clustered at the sub-district level in parentheses. All specifications include a time trend. The expenditure in levels are in 1,000 IDR. 10,000 IDR is approximately one US dollar. * $p < 0.10$, ** $p < 0.05$, ***

4.B Overview of consumption categories

The following consumption caegories (taken from Heffetz (2011) and extended by three new categories were used in my survey. The version of this survey in Bahasa Indonesia that I used to collect this data is available upon request. Also the original filled in surveys are available upon request.

Consumption Categories taken from Heffetz:

FdH: Food and nonalcoholic beverages at grocery, specialty, and convenience stores;

FdO: Dining out at restaurants, drive-throughs, etc., excluding alcohol; including food at school;

Cig: Tobacco products like cigarettes, cigars, and pipe tobacco;

AlH: Alcoholic beverages for home use;

AlO: Alcoholic beverages at restaurants, bars, cafeterias, cafe's, etc.;

Clo: Clothing and shoes, not including underwear, undergarments, and nightwear;

Lry: Laundry and dry cleaning; Jwl: Jewelry and watches;

Brb: Barbershops, beauty parlors, hair dressers, health clubs, etc.;

Hom: Rent, or mortgage, or purchase, of their housing;

Htl: Lodging away from home on trips and housing for someone away at school;

Fur: Home furnishings and household items, like furniture, appliances, tools, and linen;

Utl: Home utilities such as electricity, gas, and water; garbage collection;

Tel: Home telephone services, not including mobile phones;

Cel: Mobile phone services;

HIn: Homeowner's insurance, fire insurance, and property insurance;

Med: Medical care, including health insurance, drugs, dentists, doctors, hospitals, etc.;

Fee: Legal fees, accounting fees, and occupational expenses like tools and licenses;

LIn: Life insurance, endowment, annuities, and other death benefits insurance;

Car: The purchase of new and used motor vehicles such as cars, trucks, and vans;

CMn: Vehicle maintenance, mechanical and electrical repair and replacement;

Gas: Gasoline and diesel fuel for motor vehicles;

CIn: Vehicle insurance, like insurance for cars, trucks, and vans;

Bus: Public transportation, both local and long distance, like buses and trains;

Air: Airline fares for out-of-town trips;

Bks: Books, including school books, newspapers and magazines, toys, games, and hobbies;

Ot1: Computers, games, TVs, video, audio, musical and sports equipment, tapes, CDs;

Ot2: Cable TV, pets and veterinarians, sports, country clubs, movies, and concerts;

Education: Education, from nursery to college, like tuition and other school expenses;

CHA: Contributions to churches or other religious organisations, and other charities;

Added Consumption Categories

Pil: Pilgrimage, e.g. expenses for going to Haji;

Wed: Weddings and Feasts.

Dropped Consumption Categories:

Und: Underwear, undergarments, nightwear, and sleeping garments;

Htl: Lodging away from home on trips and housing for someone away at school.

Conclusion

Summary This thesis provides evidence that people's beliefs and their behaviour strongly change in response to new information. In Chapter 1, we show that people's beliefs about the labour market impact of immigration change after exposure to research evidence. We also provide evidence that this information strongly changes people's immigration preferences and their political behaviour. In Chapter 2, we examine whether people's beliefs about the experimenter's objectives can shape experimental participants' behaviour in experiments. Chapter 3 provides evidence that people's beliefs about government debt change in response to exposure to official statistics about government debt. Information about government debt also strongly affects people's self-reported attitudes towards government spending, and also increases people's willingness to donate to an NGO in favour of government spending cuts. Finally, in Chapter 4, we show that people's consumption choices respond to changes in their neighbours' income. These changes depend on people's social environment.

Future research We have established that people's beliefs and their behaviour respond to information. However, there are several open questions in this literature. First, it is entirely unclear under what circumstances people respond to information and which characteristics of the information matter for its effectiveness. Second, little empirical evidence exists on people's demand for information. Third, we have very little understanding of

how people allocate their attention to different types of information. Fourth, it is not clear whether results from online surveys and lab experiments carry over to natural field settings.

I have a clear-cut research agenda for the next years laid out which seeks to address some of the questions mentioned above. Specifically, my collaborators and I plan to provide evidence on both (i) the formation of beliefs, and (ii) the causal effect of beliefs on economic behaviour in the field.

In the first strand of future work, I want to combine insights from the cognitive psychology and memory research to better understand what types of information persistently change beliefs, and economic behavior. In this strand of research my collaborators and I plan to understand how different cognitive channels can explain why certain types of information (e.g. stories vs statistics; qualitative vs. quantitative information) with different characteristics (salient, non-salient) shape beliefs more persistently than others. Specifically, we plan to focus on the role of memory in shaping people's beliefs about important facts. On top of that, my collaborators and I want to study the role of motivated cognition and self-deception for belief formation. We plan to study the psychological determinants of people's demand for information from biased information sources.

In the second strand of future work, we plan to conduct large-scale natural field experiments studying the impact of beliefs on economic and political behaviour in the field. This continues the work with my collaborators studying the role of beliefs in shaping people's political engagement (Hensel et al., 2018). In this strand of research, we examined how beliefs about the effort choices of fellow party activists affect party activists' own effort choices. Specifically, we conducted a natural field experiment with a major

European party during a recent high-stakes election. We randomly provided canvassers with true information about the canvassing intentions of fellow party members. Using survey evidence and unobtrusive, geo-coded behavioural data from the party's canvassing app, we find that treated canvassers reduce their own canvassing significantly when learning that their peers engage in more canvassing than previously thought. Treatment effects are particularly large i) along the intensive margin; ii) in the final days of the campaign; iii) in competitive districts; and iv) among canvassers without prior canvassing experience. The evidence implies that effort choices of political activists exhibit strategic substitutability, consistent with seminal evidence on protest behaviour (Cantoni et al., 2017).

Future work should examine how people's economic behaviour responds to information in different field contexts. For example, it would be interesting to conduct experiments on Facebook and Twitter to gather more field evidence on the causal effect of information on political behaviour in relevant natural settings. Moreover, it seems promising to link information experiments with administrative data on consumer and financial behavior. For example, a promising avenue for future research would be to analyze how people's expectations about economic variables (such as GDP, and inflation) causally affect real economic, and financial behaviour as measured in the administrative data.

References

Agarwal, S., V. Mikhed, and B. Scholnick (2016). Does Inequality Cause Financial Distress? Evidence from Lottery Winners and Neighboring Bankruptcies. *Working Paper*.

Al-Ubaydli, O., J. A. List, D. LoRe, and D. Suskind (2017). Scaling for Economists: Lessons from the Non-adherence Problem in the Medical Literature. *Journal of Economic Perspectives 31*(4), 125–44.

Alan, S. and S. Ertac (2017). Belief in Hard Work and Prosocial Behavior: Evidence from a Randomized Field Experiment.

Alatas, V. (2011). Program Keluarga Harapan: Impact Evaluation of Indonesia's Pilot Household Conditional Cash Transfer Program. *World Bank Report*.

Alesina, A., D. Carloni, and G. Lecce (2012). The Electoral Consequences of Large Fiscal Adjustments. In *Fiscal Policy after the Financial Crisis*, pp. 531–570. University of Chicago Press.

Alesina, A. and P. Giuliano (2011). Preferences for redistribution. In *Handbook of Social Economics*, Volume 1, pp. 93–131. Elsevier.

Alesina, A. and E. La Ferrara (2000). Participation in Heterogeneous Communities. *Quarterly Journal of Economics*, 847–904.

Alesina, A. and E. La Ferrara (2005). Preferences for Redistribution in the Land of Opportunities. *Journal of Public Economics 89*(5), 897–931.

Alesina, A. and A. Passalacqua (2015). The Political Economy of Government Debt. Working Paper 21821, National Bureau of Economic Research.

Alesina, A., R. Perotti, J. Tavares, M. Obstfeld, and B. Eichengreen (1998). The Political Economy of Fiscal Adjustments. *Brookings Papers on Economic Activity 1998*(1), 197–266.

Alesina, A., S. Stantcheva, and E. Teso (2018). Intergenerational Mobility and Preferences for Redistribution. *American Economic Review 108*(2), 521–554.

Alesina, A. and G. Tabellini (1990). A Positive Theory of Fiscal Deficits and Government Debt. *The Review of Economic Studies 57*(3), 403–414.

Allcott, H. and D. Taubinsky (2015). Evaluating Behaviorally-motivated Policy: Experimental Evidence from the Lightbulb Market. *The American Economic Review 105*(8), 2501–38.

Allers, M., J. De Haan, and F. De Kam (1998). Using Survey Data to Test for Ricardian Equivalence. *Public Finance Review 26*(6), 565–582.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Andreoni, J. and C. Sprenger (2012). Estimating Time Preferences from Convex Budgets. *The American Economic Review 102*(7), 3333–56.

Angelucci, M. and G. De Giorgi (2009). Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption? *American Economic Review*, 486–508.

Angrist, J. and G. Imbens (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–476.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Ansolabehere, S., M. Meredith, and E. Snowberg (2013). Asking about Numbers: Why and how. *Political Analysis 21*(1), 48–69.

Armona, L. C., A. Fuster, and B. Zafar (2018). Home Price Expectations and Behavior: Evidence from a Randomized Information Experiment. *Review of Economic Studies (forthcoming)*.

Athey, S. and G. Imbens (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Bagwell, L. S. and B. D. Bernheim (1996). Veblen Effects in a Theory of Conspicuous Consumption. *American Economic Review*, 349–373.

Bandiera, O., I. Barankay, and I. Rasul (2009). Social Connections and Incentives in the Workplace: Evidence from Personnel Data. *Econometrica 77*(4), 1047–1094.

Bardsley, N. (2008). Dictator Game Giving: Altruism or Artefact? *Experimental Economics 11*(2), 122–133.

Barmettler, F., E. Fehr, and C. Zehnder (2012). Big Experimenter is Watching you! Anonymity and Prosocial Behavior in the Laboratory. *Games and Economic Behavior 75*(1), 17–34.

Barro, R. J. (1974). Are Government Bonds Net Wealth? *Journal of Political Economy 82*(6), 1095–1117.

Battaglini, M. and S. Coate (2008). A Dynamic Theory of Public Spending, Taxation, and Debt. *The American Economic Review 98*(1), 201–236.

Battaglini, M., S. Nunnari, and T. R. Palfrey (2016). The Political Economy of Public Debt: A Laboratory Study. Working Paper 22406, National Bureau of Economic Research.

Becker, S. O. and T. Fetzer (2016). Does Migration Cause Extreme Voting. Technical report, Competitive Advantage in the Global Economy (CAGE).

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior 10*(1), 122–142.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis 20*(3), 351–368.

Berinsky, A. J., M. F. Margolis, and M. W. Sances (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science 58*(3), 739–753.

Bertrand, M., E. F. Luttmer, and S. Mullainathan (2000). Network Effects and Welfare Cultures. *The Quarterly Journal of Economics 115*(3), 1019–1055.

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

Bertrand, M. and S. Mullainathan (2001). Do People Mean What They Say? Implications for Subjective Survey Data. *The American Economic Review 91*(2), 67–72.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 992–1026.

Binmore, K., A. Shaked, and J. Sutton (1985). Testing Noncooperative Bargaining Theory: A Preliminary Study. *The American Economic Review 75*(5), 1178–1180.

Bischoff, I. and B. Frank (2011). Good News for Experimenters: Subjects are Hard to Influence by Instructors' Cues. *Economics Bulletin 31*(4), 3221–3225.

Bolton, G. E., E. Katok, and R. Zwick (1998). Dictator Game Giving: Rules of Fairness Versus Acts of Kindness. *International Journal of Game Theory 27*(2), 269–299.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Beliefs about Gender. *Mimeo, Harvard University*.

Borjas, G. J. (2014). *Immigration Economics*. Harvard University Press.

Borjas, G. J. (2017). The Wage Impact of the Marielitos: A Reappraisal. *ILR Review 70*, 1077–1110.

Brender, A. and A. Drazen (2008). How Do Budget Deficits and Economic Growth Affect Reelection Prospects? Evidence from a Large Panel of Countries. *The American Economic Review 98*(5), 2203–2220.

Bursztyn, L., M. Callen, B. Ferman, S. Gulzar, A. Hasanain, and N. Yuchtman (2016). Political Identity: Experimental Evidence on Anti-americanism in Pakistan. *Working Paper*.

Bursztyn, L., F. Ederer, B. Ferman, and N. Yuchtman (2014). Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions. *Econometrica 82*(4), 1273–1301.

Bursztyn, L., B. Ferman, S. Fiorin, M. Kanz, and G. Rao (2018). Status Goods: Experimental Evidence from Platinum Credit Cards. *Quarterly Journal of Economics (forthcoming)*.

Bursztyn, L. and R. Jensen (2015). How Does Peer Pressure Affect Educational Investments? *The Quarterly Journal of Economics 130*(3), 1329–1367.

Camerer, C. F. (2012). The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List. In *The Methods of Modern Experimental Economics*. Oxford University Press.

Camerer, C. F., R. M. Hogarth, D. V. Budescu, and C. Eckel (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-production framework. *Journal of Risk and Uncertainty*, 7–48.

Cameron, A. C. and D. L. Miller (2015). A Practitioner's Guide to Cluster-robust Inference. *Journal of Human Resources 50*(2), 317–372.

Cantoni, D., D. Y. Yang, N. Yuchtman, and Y. J. Zhang (2017). Protests as Strategic

Games: Experimental Evidence from Hong Kong?s Democracy Movement. *Working Paper*.

Card, D. (1990). The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relations Review 43*(2), 245–257.

Card, D. (2005). Is the New Immigration Really so Bad? *Economic Journal 115*(507).

Card, D. (2012). Comment: The Elusive Search for Negative Wage Impacts of Immigration. *Journal of the European Economic Association 10*(1), 211–215.

Card, D., C. Dustmann, and I. Preston (2012). Immigration, Wages, and Compositional Amenities. *Journal of the European Economic Association 10*(1), 78–119.

Card, D., A. Mas, E. Moretti, and E. Saez (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review 102*(6), 2981–3003.

Card, D. and G. Peri (2016). *Immigration Economics* by George J. Borjas: A Review Essay. *Journal of Economic Literature 54*(4), 1333–1349.

Cavallo, A., G. Cruces, and R. Perez-Truglia (2016). Inflation Expectations, Learning and Supermarket Prices: Evidence from Field Experiments. *American Economic Journal: Macroeconomics 9*(3), 1–35.

Charles, K. K., E. Hurst, and N. Roussanov (2009). Conspicuous Consumption and Race. *Quarterly Journal of Economics 124*(2), 425–467.

Charness, G., U. Gneezy, and M. A. Kuhn (2012). Experimental Methods: Between-subject and Within-subject Design. *Journal of Economic Behavior & Organization 81*(1), 1–8.

Chassang, S., G. Padró i Miquel, and E. Snowberg (2012). Selective Trials: A Principal-agent Approach to Randomized Controlled Experiments. *The American Economic Review 102*(4), 1279–1309.

Chudik, A., K. Mohaddes, M. H. Pesaran, and M. Raissi (2016). Is There a Debt-Threshold Effect on Output Growth? *Review of Economics and Statistics 99*(1), 135–150.

Cilliers, J., O. Dube, and B. Siddiqi (2015). The White-man Effect: How Foreigner Presence Affects Behavior in Experiments. *Journal of Economic Behavior & Organization 118*, 397–414.

Citrin, J., D. P. Green, C. Muste, and C. Wong (1997). Public Opinion Toward Immigration Reform: The Role of Economic Motivations. *Journal of Politics 59*(3), 858–881.

Clark, H. H. and M. F. Schober (1992). Asking Questions and Influencing Answers. In J. M. Tanur (Ed.), *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*. Russell Sage Foundation.

Clemens, M. A. (2017). What the Mariel Boatlift of Cuban Refugees Can Teach Us about the Economics of Immigration: An Explainer and a Revelation.

Clemens, M. A. and J. Hunt (2017). The Labor Market Effects of Refugee Waves: Reconciling Conflicting Results. Working Paper 23433, National Bureau of Economic Research.

Coffman, L., C. R. Featherstone, and J. B. Kessler (2015). A Model of Information Nudges. *Working paper*.

Croson, R. and U. Gneezy (2009). Gender Differences in Preferences. *Journal of Economic Literature 47*(2), 448–474.

Cruces, G., R. Perez-Truglia, and M. Tetaz (2013). Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment. *Journal of Public Economics 98*, 100–112.

Cukierman, A. and A. H. Meltzer (1989). A Political Theory of Government Debt and Deficits in a Neo-Ricardian Framework. *The American Economic Review*, 713–732.

Dahl, G. B., K. V. Løken, and M. Mogstad (2014). Peer Effects in Program Participation. *American Economic Review 104*(7), 2049–74.

Dal Bó, E. and P. Dal Bó (2014). "Do the Right Thing:" The Effects of Moral Suasion on Cooperation. *Journal of Public Economics 117*, 28–38.

De Giorgi, G., A. Frederiksen, and L. Pistaferri (2016). Consumption Network Effects. *NBER working paper*.

de Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and Bounding Experimenter Demand. *American Economic Review* (forthcoming).

de Quidt, J., L. Vesterlund, and A. Wilson (2018). Experimenter Demand Effects. In A. Ule and A. Schram (Eds.), *Handbook of Research Methods and Applications in Experimental Economics, forthcoming*. Edward Elgar Publishing.

Dehdari, S. (2017). Economic Distress and Support for Far-right Parties - Evidence from Sweden. *Job Market Paper*.

DellaVigna, S., J. A. List, and U. Malmendier (2012). Testing for Altruism and Social Pressure in Charitable Giving. *The Quarterly Journal of Economics 127*, 1–56.

DellaVigna, S., J. A. List, U. Malmendier, and G. Rao (2017). Voting to Tell Others. *Review of Economic Studies 84*(1), 143–181.

DellaVigna, S. and D. Pope (2017). Predicting Experimental Results: Who Knows What? *Journal of Political Economy, forthcoming*.

DellaVigna, S. and D. Pope (2018). What Motivates Effort? Evidence and Expert Forecasts. *Review of Economic Studies 85*(2), 1029–1069.

Downs, A. (1957). An Economic Theory of Political Action in a Democracy. *Journal of Political Economy 65*(2), 135–150.

Duflo, E. and E. Saez (2003). The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *Quarterly Journal of Economics 118*(3), 815–842.

Dupas, P. and E. Miguel (2017). Impacts and Determinants of Health Levels in Low-income Countries. In *Handbook of Economic Field Experiments*, Volume 2, pp. 3–93. Elsevier.

Dustmann, C., F. Fabbri, and I. Preston (2005). The Impact of Immigration on the British Labour Market. *Economic Journal 115*(507), 324–341.

Dustmann, C. and I. Preston (2001). Attitudes to Ethnic Minorities, Ethnic Context and Location Decisions. *Economic Journal 111*(470), 353–373.

Dustmann, C. and I. Preston (2006). Is Immigration Good or Bad for the Economy? Analysis of Attitudinal Responses. *Research in Labor Economics 24*(1), 3–34.

Dustmann, C., K. Vasiljeva, and A. P. Damm (2018). Refugee Migration and Electoral Outcomes. *Review of Economic Studies (forthcoming)*.

Ellingsen, T., R. Östling, and E. Wengström (2018). How Does Communication Affect Beliefs in One-shot Games With Complete Information? *Games and Economic Behavior 107*, 153–181.

Facchini, G., Y. Margalit, and H. Nakata (2016). Countering Public Opposition to Immigration: The Impact of Information Campaigns. *Working Paper*.

Facchini, G., A. M. Mayda, and R. Puglisi (2009). Illegal immigration and media exposure: Evidence on individual attitudes. *CEPR Discussion Papers 7593*.

Falk, A. and J. J. Heckman (2009). Lab Experiments are a Major Source of Knowledge in the Social Sciences. *Science 326*(5952), 535–538.

Falk, A. and F. Zimmermann (2012). A Taste for Consistency and Survey Response Behavior. *CESifo Economic Studies 59*(1), 181–193.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in Disguise—an Experimental Study on Cheating. *Journal of the European Economic Association 11*(3), 525–547.

Fiszbein, A., N. R. Schady, and F. H. Ferreira (2009). *Conditional Cash Transfers: Reducing Present and Future Poverty*. World Bank Publications.

Fleming, P. and D. J. Zizzo (2014). A Simple Stress Test of Experimenter Demand Effects. *Theory and Decision 78*(2), 219–231.

Frank, R. H. (1985). The Demand for Unobservable and Other Nonpositional Goods. *The American Economic Review 75*(1), 101–116.

Gerber, A., M. Hoffman, J. Morgan, and C. Raymond (2017). One in a Million: Field Experiments on Perceived Closeness of the Election and Voter Turnout. Technical report, National Bureau of Economic Research.

Gilens, M. (2001). Political Ignorance and Collective Policy Preferences. *American Political Science Review 95*(2), 379–396.

Giuliano, P. and A. Spilimbergo (2014). Growing up in a Recession. *Review of Economic Studies 81*(2), 787–817.

Glazer, A. and K. A. Konrad (1996). A Signaling Explanation for Charity. *The American Economic Review 86*(4), 1019–1028.

Gneezy, U. and J. Potters (1997). An Experiment on Risk Taking and Evaluation Periods. *The Quarterly Journal of Economics*, 631–645.

Grigorieff, A., C. Roth, and D. Ubfal (2016). Does Information Change Attitudes Towards Immigrants? Representative Evidence From Survey Experiments. *IZA Discussion Paper No. 10419*.

Grinblatt, M., M. Keloharju, and S. Ikäheimo (2008). Social Influence and Consumption: Evidence from the Automobile Purchases of Neighbors. *Review of Economics and Statistics 90*(4), 735–753.

Güth, W., R. Schmittberger, and B. Schwarze (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization 3*(4), 367–388.

Haaland, I. and C. Roth (2017). Labor Market Concerns and Support for Immigration. Available at SSRN: https://ssrn.com/abstract=3017420.

Hainmueller, J. and M. J. Hiscox (2010). Attitudes Toward Highly Skilled and Low-skilled Immigration: Evidence from a Survey Experiment. *American Political Science Review 104*(01), 61–84.

Hainmueller, J., M. J. Hiscox, and Y. Margalit (2015). Do Concerns about Labor Market Competition Shape Attitudes Toward Immigration? New Evidence. *Journal of International Economics 97*(1), 193–207.

Hainmueller, J. and D. J. Hopkins (2014a). The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes Toward Immigrants. *American Journal of Political Science 59*(3), 529–548.

Hainmueller, J. and D. J. Hopkins (2014b). Public Attitudes Toward Immigration. *Annual Review of Political Science 17*, 225–249.

Halla, M., A. F. Wagner, and J. Zweimüller (2017). Immigration and Voting for the Far Right. *Journal of the European Economic Association 15*(6), 1341–1385.

Harrison, G. W. and J. A. List (2004). Field Experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Hauser, D. J. and N. Schwarz (2016). Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than do Subject Pool Participants. *Behavior research methods 48*(1), 400–7.

Hayo, B. and F. Neumeier (2016). The (In-)Validity of the Ricardian Equivalence Theorem: Findings from a Representative German Population Survey.

Heffetz, O. (2011). A Test of Conspicuous Consumption: Visibility and Income Elasticities. *Review of Economics and Statistics 93*(4), 1101–1117.

Heffetz, O. and R. H. Frank (2008). Preferences for Status: Evidence and Economic Implications. *Handbook of social economics 1*, 69–91.

Heinemann, F. and T. Hennighausen (2012). Understanding Public Debt Preferences. *FinanzArchiv: Public Finance Analysis 68*(4), 406–430.

Hensel, L., J. Hermle, A. Rink, and C. Roth (2018). Political Activists as Free-Riders: Evidence From a Natural Field Experiment. *Working Paper*.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith (1994). Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior 7*(3), 346–380.

Hoffman, E., K. A. McCabe, and V. L. Smith (1996). On Expectations and the Monetary Stakes in Ultimatum Games. *International Journal of Game Theory 25*(3), 289–301.

Hopkins, D. J., J. Sides, and J. Citrin (2018). The Muted Consequences of Correct Information About Immigration. *Journal of Politics (forthcoming)*.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica: Journal of the Econometric Society*, 467–475.

Imbens, G. W. and C. F. Manski (2004). Confidence Intervals for Partially Identified Parameters. *Econometrica 72*(6), 1845–1857.

Ireland, N. J. (1994). On Limiting the Market for Status Signals. *Journal of Public Economics 53*(1), 91–110.

Iyengar, S., S. Jackman, S. Messing, N. Valentino, T. Aalberg, R. Duch, K. S. Hahn, S. Soroka, A. Harell, and T. Kobayashi (2013). Do attitudes about immigration predict willingness to admit individual immigrants? A cross-national test of the person-positivity bias. *Public Opinion Quarterly 77*(3), 641–665.

Jakiela, P. and O. W. Ozier (2016). Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies. *Review of Economic Studies 83*(1), 231–268.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986). Fairness and the Assumptions of Economics. *Journal of Business*, S285–S300.

Karadja, M., J. Mollerstrom, and D. Seim (2017). Richer (and Holier) than Thou? The Effect of Relative Income Improvements on Demand for Redistribution. *Review of Economics and Statistics 99*(2), 201–212.

Kaus, W. (2013). Conspicuous Consumption and Race: Evidence from South Africa. *Journal of Development Economics 100*(1), 63–73.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental Analysis of Neighborhood Effects. *Econometrica 75*(1), 83–119.

Kuhn, P., P. Kooreman, A. Soetevent, and A. Kapteyn (2011). The Effects of Lottery Prizes on Winners and their Neighbors: Evidence from the Dutch Postcode Lottery. *American Economic Review 101*(5), 2226–2247.

Kuklinski, J. H., P. J. Quirk, J. Jerit, D. Schwieder, and R. F. Rich (2000). Misinformation and the Currency of Democratic Citizenship. *Journal of Politics 62*(3), 790–816.

Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015). How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments. *American Economic Review 105*(4), 1478–1508.

Lambdin, C. and V. A. Shaffer (2009). Are Within-subjects Designs Transparent? *Judgment and Decision Making 4*(7), 554–566.

Lergetporer, P., M. Piopiunik, L. Simon, et al. (2017). Do Natives' Beliefs About Refugees' Education Level Affect Attitudes Toward Refugees? Evidence from Randomized Survey Experiments. Technical report, CESifo Group Munich.

Lergetporer, P., G. Schwerdt, K. Werner, and L. Woessmann (2016). Information and Preferences for Public Spending: Evidence from Representative Survey Experiments. *IZA Discussion Paper No. 9968*.

Levitt, S. D. and J. A. List (2007). What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *The Journal of Economic Perspectives 21*(2), 153–174.

List, J. A. (2006). The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy 114*(1), 1–37.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy 115*(3), 482–493.

List, J. A., R. P. Berrens, A. K. Bohara, and J. Kerkvliet (2004). Examining the Role of Social Isolation on Stated Preferences. *The American Economic Review 94*(3), 741–752.

Loewenstein, G. (1999). Experimental Economics From the Vantage-point of Behavioural Economics. *The Economic Journal 109*(453), 25–34.

Luttmer, E. F. (2005). Neighbors as Negatives: Relative Earnings and Well-Being. *Quarterly Journal of Economics 120*(3), 963–1002.

Manacorda, M., A. Manning, and J. Wadsworth (2012). The Impact of Immigration on the Structure of Wages: Theory and Evidence from Britain. *Journal of the European Economic Association 10*(1), 120–151.

Manski, C. F. (1993). Identification of Endogenous Social Effects: the Reflection Problem. *Review of Economic Studies 60*(3), 531–542.

Manski, C. F. (2004). Measuring Expectations. *Econometrica 72*(5), 1329–1376.

Mas, A. and E. Moretti (2009). Peers at Work. *American Economic Review 99*(1), 112–45.

Mayda, A. M. (2006). Who is Against Immigration? A Cross-country Investigation of Individual Attitudes Toward Immigrants. *Review of Economics and Statistics 88*(3), 510–530.

Mayer, T. (1995). *Doing Economic Research*. Edward Elgar, Aldershotg.

Milgram, S. (1963). Behavioral Study of Obedience. *The Journal of Abnormal and Social Psychology 67*(4), 371.

Moav, O. and Z. Neeman (2012). Saving Rates and Poverty: The Role of Conspicuous Consumption and Human Capital. *Economic Journal 122*(563), 933–956.

Moretti, E. (2011). Social Learning and Peer Effects in Consumption: Evidence from Movie Sales. *Review of Economic Studies 78*(1), 356–393.

Müller, A., K. Storesletten, and F. Zilibotti (2016). The Political Color of Fiscal Responsibility. *Journal of the European Economic Association 14*(1), 252–302.

Mummolo, J. and E. Peterson (2017). Demand Effects in Survey Experiments: An Empirical Assessment. available at SSRN: https://ssrn.com/abstract=2956147.

Orne, M. T. (1962). On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications. *American Psychologist 17*(11), 776.

Ottaviano, G. I. and G. Peri (2012). Rethinking the Effect of Immigration on Wages. *Journal of the European Economic association 10*(1), 152–197.

Peltzman, S. (1992). Voters as Fiscal Conservatives. *The Quarterly Journal of Economics 107*(2), 327–361.

Perez-Truglia, R. (2013). A Test of the Conspicuous Consumption Model using Subjective Well-Being Data. *The Journal of Socio-Economics 45*, 146–154.

Perez-Truglia, R. (2016). The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment. *Working Paper*.

Perez-Truglia, R. and G. Cruces (2017). Partisan Interactions: Evidence from a Field Experiment in the United States. *Journal of Political Economy 125*(4), 1208–1243.

Persson, T. and L. E. Svensson (1989). Why a Stubborn Conservative would Run a Deficit: Policy with Time-Inconsistent Preferences. *The Quarterly Journal of Economics 104*(2), 325–345.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research.* Appleton-Century-Crofts.

Roth, C. P. (2014). Conspicuous Consumption and Peer Effects Among the Poor: Evidence From a Field Experiment. *CSAE Working Paper WPS/2014-29*.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics 116*(2), 681–704.

Scheve, K. F. and M. J. Slaughter (2001). Labor Market Competition and Individual Preferences over Immigration Policy. *Review of Economics and Statistics 83*(1), 133–145.

Shmaya, E. and L. Yariv (2016). Experiments on Decisions Under Uncertainty: A Theoretical Framework. *The American Economic Review 106*(7), 1775–1801.

Sides, J. and J. Citrin (2007). How Large the Huddled Masses? The Causes and Consequences of Public Misperceptions About Immigrant Populations. In *Annual Meeting of the Midwest Political Science Association, Chicago*.

Sniderman, P. M., L. Hagendoorn, and M. Prior (2004). Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities. *American Political Science Review 98*(01), 35–49.

Song, Z., K. Storesletten, and F. Zilibotti (2012). Rotten Parents and Disciplined Children: A Politico-Economic Theory of Public Expenditure and Debt. *Econometrica 80*(6), 2785–2803.

Tsutsui, K. and D. J. Zizzo (2014). Group Status, Minorities and Trust. *Experimental Economics 17*, 215–244.

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

Veblen, T. (1899). *The Theory of the Leisure Class: an Economic Study of Institutions*. London: Macmillan Publishers.

West, M. R., L. Woessmann, P. Lergetporer, and K. Werner (2016). How Information Affects Support for Education Spending: Evidence from Survey Experiments in Germany and the United States. *CESifo Working Paper*.

Zizzo, D. J. (2010). Experimenter Demand Effects in Economic Experiments. *Experimental Economics 13*(1), 75–98.